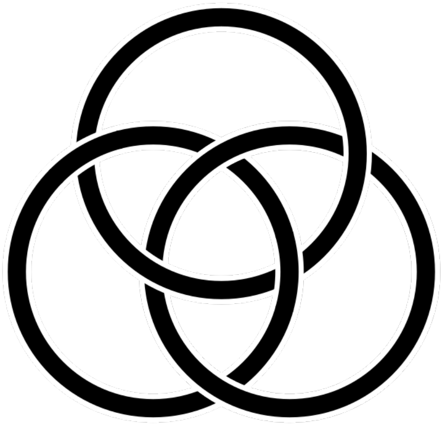
Higher-order interactions in single-cell gene expression

Towards a cybergenetic semantics of cell state

Abel Jansma

Institute of Genetics & Cancer
Higgs Centre for Theoretical Physics



A thesis submitted for the degree of
Doctor of Philosophy

June 14, 2023

MRC Human
Genetics
Unit

THE UNIVERSITY
of EDINBURGH

Abstract

Finding and understanding patterns in gene expression guides our understanding of living organisms, their development, and diseases, but is a challenging and high-dimensional problem as there are many molecules involved. One way to learn about the structure of a gene regulatory network is by studying the interdependencies among its constituents in transcriptomic data sets. These interdependencies could be arbitrarily complex, but almost all current models of gene regulation contain pairwise interactions only, despite experimental evidence existing for higher-order regulation that cannot be decomposed into pairwise mechanisms. I set out to capture these higher-order dependencies in single-cell RNA-seq data using two different approaches. First, I fitted maximum entropy (or Ising) models to expression data by training restricted Boltzmann machines (RBMs). On simulated data, RBMs faithfully reproduced both pairwise and third-order interactions. I then trained RBMs on 37 genes from a scRNA-seq data set of 70k astrocytes from an embryonic mouse. While pairwise and third-order interactions were revealed, the estimates contained a strong omitted variable bias, and there was no statistically sound and tractable way to quantify the uncertainty in the estimates. As a result I next adopted a model-free approach. Estimating model-free interactions (MFIs) in single-cell gene expression data required a quasi-causal graph of conditional dependencies among the genes, which I inferred with an MCMC graph-optimisation algorithm on an initial estimate found by the Peter-Clark algorithm. As the estimates are model-free, MFIs can be interpreted either as mechanistic relationships between the genes, or as substructures in the cell population. On simulated data, MFIs revealed synergy and higher-order mechanisms in various logical and causal dynamics more accurately than any correlation- or information-based quantities. I then estimated MFIs among 1,000 genes, at up to seventh-order, in 20k neurons and 20k astrocytes from two different mouse brain scRNA-seq data sets: one developmental, and one adolescent. I found strong evidence for up to fifth-order interactions, and the MFIs mostly disambiguated direct from indirect regulation by preferentially coupling causally connected genes, whereas correlations persisted across causal chains. Validating the predicted interactions against the Pathway Commons database, gene ontology annotations, and semantic similarity, I found that pairwise MFIs contained different but a similar amount of mechanistic information relative to networks based on correlation. Furthermore, third-order interactions provided evidence of combinatorial regulation by transcription factors and immediate early genes.

I then switched focus from mechanism to population structure. Each significant MFI can be assigned a set of single cells that most influence its value. Hierarchical clustering of the MFIs by cell assignment revealed substructures in the cell population corresponding to diverse cell states. This offered a new, purely data-driven view on cell states because the inferred states are not required to localise in gene expression space. Across the four data sets, I found 69 significant and biologically interpretable cell states, where only 9 could be obtained by standard approaches. I identified immature neurons among developing astrocytes and radial glial cells, D1 and D2 medium spiny neurons, D1 MSN subtypes, and cell-cycle related states present across four data sets. I further found evidence for states defined by genes associated to neuropeptide signalling, neuronal activity, myelin metabolism, and genomic imprinting. MFIs thus provide a new, statistically sound method to detect substructure in single-cell gene expression data, identifying cell types, subtypes, or states that can be delocalised in gene expression space and whose hierarchical structure provides a new view on the semantics of cell state. The estimation of the quasi-causal graph, the MFIs, and inference of the associated states is implemented as a publicly available Nextflow pipeline called Stator.

To my family

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where stated otherwise by reference or acknowledgment, the work presented is entirely my own.

Parts of this thesis, in particular the results from Chapter 3, have previously appeared as a preprint *Higher-order in-and-outeractions reveal synergy and logical dependence beyond Shannon-information - A. Jansma 2022*, available at https://arxiv.org/abs/2205.04440.

Abel Jansma – Berlin, June 14, 2023

Lay Summary

Biology and its cast members Biology is the study of one of the biggest mysteries of our universe: Life. There are many deep questions to ask about life: What is Life? Where does Life come from? How does Life begin, persist, and end? Like all sciences, biology aims to answer such deep question by figuring out how things work. To describe how *a Thing* works, you need to describe the components that make up *the Thing*, how these components come together to form *the Thing*, and how these components interact. Perhaps surprisingly, it is often already difficult to come up with the right components to base your descriptions of Nature on. In the same way that a written play often starts with the list of characters that will appear in the coming scenes, any story about Nature should start with a meaningful list of interacting parts—the cast. Physicists have it easy, because there are fundamental *things*—called elementary particles—that make up every*thing*. Biologists face an especially daunting task, since there seem to be interesting characters at *every* scale of Life: Enormous ecosystems are made from interacting species, which are composed of individual organisms, which are made from cells and other organisms living inside them, each of which has echoes of its ancestors and their experiences scattered throughout its genetic material and might host many viruses. None of these levels unequivocally deserves to be called *fundamental*, so any description of life depends on which list of characters it is based on. Plato recognised this choice as the central challenge of what we now call science, and called it *Carving Nature at its joints*.

The atlases of life The importance of decomposing wholes into parts has filled much of the history of biology with atlases—catalogues of the wide variety of cast members—of the various ecologies on earth and interconnected webs of species and food chains, to the diversity of seashells and beetles. Many scientific breakthroughs have come from describing and understanding the structure of these catalogues of life. Most famously, Darwin described the source of all species on earth and their relationships in *On the Origin of Species*.

Modern biology has found one particularly interesting collection of entities that can explain some of the diversity in life: Genes. Genes are interesting because while they are physically small, they are present in all living things, and reveal the entangled history of life on earth. They describe not only the diversity of species on earth, but also the variability within a species. However, they leave one final source of diversity unexplained: why do the cells that make up an individual—that make up *you*—all look different, even though they contain the same DNA? That is the question this thesis is based on. I wanted to describe the differences between cells by describing how genes interact and come together to constitute the cell's identity. This is important because it would describe how our bodies develop from an embryo into an adult, how cells respond to each other and the environment, but could also provide insight into diseases, many of which begin as a problem in one or more cells.

Listening in on genetic conversations One description of cell biology is so famous and ubiquitous, it is known as the *central dogma*. It describes how genes (DNA molecules) are used to make proteins that perform functions in the cell. This alone is not enough to make cells with identical DNA behave differently, but it turns out that the proteins that are made from the DNA can interact with the DNA itself to change which other proteins get made. By doing this, cells can control what happens inside them and change their behaviour.

To gain insight into these regulatory relationships, I 'listened in' on the way the genes were active in the cells. This is possible, because each time a gene is used by the cell, it produces a single molecule, called an RNA transcript, that is unique to that gene. By counting the transcripts in a single cell, you can get an idea of which genes were active, also called *expressed*. If you do this in many cells, and see for example that two genes are always active together, or inactive together, then these genes might be coordinating their expression. Doing this for all genes can give a description of which genes coordinate their expression together. This has been done many times before, and is a common technique throughout biology. However, such descriptions currently only find *pairs* of genes that *interact*, that is, two genes that coordinate their expression. However, there might be more complex coordination present among the genes. What does coordination between three genes look like? Which genes do this? What does such coordination lead to? These are the first questions I tried to answer.

I found that many genes do coordinate their expression in such *higher-order* interactions, involving up to five genes at the same time. The genes that did so tended to be special kinds of genes called *transcription factors*. These are genes whose main function is to regulate the expression of other genes. It was already known that they can do this in complex ways, regulating the expression of many other genes in various ways, and the fact that precisely these genes show these higher-order interactions supports this idea.

Given a set of genes with such a higher-order interaction, I asked: in which cells do these genes most strongly coordinate their expression? This set of cells corresponded to the cells in which the genes are collaborating to get something done. Like this, I found interactions that revealed which cells were in the process of dividing, for example. I call such a set of cells with conspiring genes a *cell state*. I found cell states that relate to active neurons, states with cells that were slowly transforming into other kinds of cells, states that corresponded to special kind of neurons not visible with other techniques, and many more.

Together, these results showed that such higher-order interactions are indeed important to descriptions of gene expression, and that they can reveal some of the diversity among cells that results from gene expression.

Acknowledgements

> Let me keep company always with
> those who say "Look!" and laugh in
> astonishment, and bow their heads.
>
> Mary Oliver [183]

䷌

Doing the research and writing a thesis to complete a PhD can be a challenging, long, and alienating project. This was exacerbated by the fact that the last two and a half years have seen a virus disrupt life globally. Especially in such circumstances, it becomes clear that we depend—so deeply—on those around us. This thesis is the result of many wonderful higher-order interactions I've had with those who helped and supported me during the last four years.

First of all, I want to thank my three supervisors: Chris, Luigi, and Ava. Your warm and generous guidance has taught me so much about the ups and downs of research, collaboration, and Nature, and I deeply appreciate the freedom you have given me to pursue the many distractions I encountered and created. Thanks also to my thesis committee—Martin Taylor, Catalina Vallejos, Paul McKeigue, and Ailith Ewing—who have provided valuable feedback and discussions during my annual reviews, and to my two examiners Matthias Hennig and Caleb Webber for their time and interest in my work. For the many discussions about causality, interactions and abstract nonsense, I would like to thank Ava and Sjoerd.

I was lucky enough to be embedded in the Ponting lab at the Institute of Genetics and Cancer, and was warmly welcomed by all its members. In particular, I would like to thank Øyvind for sharing his experience with software development, for all the rubber ducking, and for being so generous with his time. Thanks also to Tilly and Josh, who kept me company during many talks and coffee breaks at the IGC. My gratitude also goes out to Yuelin, Jareth, and Richard, who patiently worked with unstable versions of the pipeline and discovered many bugs.

Even beyond Edinburgh, many scientists took the time to share their knowledge with me, and I want to thank them all. In particular, I want to express how valuable my conversations with Gemma de las Cuevas have been. Exploring and discussing our shared universal curiosity has deepened my appreciation and understanding of the mystery of Life.

I had the privilege of moving to Edinburgh with Miró, who became a dear and close friend during our time living together, as did Hannah and Kees, with whom I spent many nights exploring Leith and the rest of Scotland. My second year in Scotland was spent with the Cumings family, whose warmth, curiosity, openness, and generosity have deeply moved me, and always made me feel welcome. Thank you so much, Sarah, Charles, Mercy, Jesse, Eve, Eartha, Jonah, and everyone around you.

I want to emphasise that I could not have done this without those closest to me. First of all, my family, in whom I could always find a safe and loving home during these tumultuous years. Each one of you has supported and inspired me in their own way. I want to thank Rein and Maartje, for showing me everything; Noa, for questioning everything; and Mas, for loving everything.

Finally, my love, Julia. For many years, your love stretched along the Rhine and across the North Sea, and always supported me. Two years ago, we finally found a home together, and I feel incredibly lucky to share my life with you. It has been an infinite adventure—I can't wait for what is to come.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> Gradually, a general connection presents itself—not a linear one, but a net-like entangled fabric, with higher-order formation and destruction, with many fluctuations in the relative proportions of the parts—through the curious inquiry of Nature.
>
> Alexander von Humboldt [286], p.33 (translation mine)

䷂

1.1 The importance of stamp collecting

> Philately starts where the catalogue ends.
>
> Anonymous

Biology is often called a *messy* science, full of noise, exceptions, and contingencies. Indeed, general but predictive biological principles are few and far between, but the ones that have been found have turned out to be very deep explanations, with an explanatory range far beyond the limits of biology. Three of the most stunning examples of general principles of life were all published in the 19th century. In 1845, Alexander von Humboldt published his *Kosmos* [286], in which he described the different ways life organises itself within various environments as an interconnected web, laying the foundations for ecology [304]. Inspired by Humboldt's ecologies, Darwin published his *On the origin of species* [63] in 1859, only fourteen years later, describing in detail his theory of evolution. Six years later, Mendel published his (initially ignored) study of hereditary traits in pea plants [162], introducing the study of genetics. Here, we have the three central principles of biology—ecology, evolution, and genetics—all published within two decades of the 19th century. Even though the principles apply to life at very different scales—ecosystems, species, and organisms, respectively—they share a common pattern. Each of the discoveries started with the observation that there is a richness and diversity to living systems that needs an explanation. Humboldt offered an explanation for the diversity in ecologies, Darwin for the wealth of species on earth, and Mendel explained the variation of traits within a species. This illustrates the importance of cataloguing the diversity of life, and stands in stark contrast to the critique, generally attributed to Ernest Rutherford, that all science is either physics or stamp collecting.

In fact, we might go further, and take the catalogues of life's diversity to be the central observations that need to be explained by biology. This leads us to the current state of biology, at a scale beyond that of even Mendel's individuals, to look at heterogeneity *within* an individual. While genetics can explain heterogeneous phenotypes among members of a species, it cannot explain why any given individual is composed of cells that each contain the same genetic material[1], but look and behave very differently. Furthermore, each cell can be traced back to a single zygote with a single genome and phenotype. It is this emergent cellular diversity in particular that is crucial to understand, since not only does it form the basis for the development and functioning of all multicellular life, it also lies at the root of most diseases, many of which seem to manifest themselves first as a cellular phenotype, before disrupting the functioning of a tissue or organism. The most striking example of this is cancer, where the disruption of a single cell's behaviour can result in unpredictable, organism-wide, and typically fatal phenotypes. The need to catalogue and understand how a single genome can lead to different cellular phenotypes is thus twofold: It is both the next obvious source of variation in need of an explanation, and it could help characterise and understand many diseases [195].

To address and describe this diversity, we need to turn our attention to molecular biology. Molecular biology is generally understood in terms of the *central dogma* [60], illustrated

[1] Modulo mutations, mosaicism, and germ-line cells.

on the left of Figure 1.1, which says that the genetic sequence information flows from DNA to RNA to proteins in processes that are known as transcription and translation, respectively. In particular, it states that sequence information cannot flow from proteins to nucleic acids. During transcription, the protein RNA-polymerase binds to a DNA molecule at a specific location along the sequence, and generates an RNA molecule, complementary to the DNA to which it is bound. This messenger-RNA (mRNA) molecule gets transported from the nucleus to the cytoplasm, where ribosomes use the RNA molecule to generate the corresponding sequence of amino acids and form it into a protein (translation).

The central dogma illustrates that if the rates at which transcription and translation happen vary across cells, then that leads to different concentrations of molecules inside the cells, potentially altering their behaviour and morphology. However, it still falls short of explaining how cells with identical DNA can show different phenotypes, as it provides no mechanism by which these rates can change. The crucial issue that the central dogma does not address is how RNA-polymerase binds to any particular location on the DNA, and when the mRNA gets translated by a particular ribosome. That is because not all information flows within a cell correspond to sequence information of either nucleic- or amino acids. There are many more processes involved in the communication and control between these three classes of molecules, as illustrated on the right of Figure 1.1, and there are important degrees of freedom besides sequence. For instance, the three classes of molecules can affect themselves: DNA can affect its own spatial configuration depending on its sequence and epigenetic modifications [311, 130], RNA molecules can bind to themselves or other RNA molecules, changing their tertiary structure or the accessibility of other RNA molecules, and proteins can bind and fold other proteins to change their behaviour. However, perhaps the most important flow of information not included in the central dogma is that of proteins to DNA. Proteins can bind to DNA, changing the DNA's three-dimensional chromatin structure and the rate at which transcription of a particular sequence can occur. Protein-DNA binding consists primarily of the formation of hydrogen bonds and van-der-Waals forces that can be highly sequence- and protein-specific. These binding reactions close the loop in Figure 1.1, which explicitly introduces *regulation* into the system, allowing the genome to regulate its own expression. The cellular dynamics are thus driven by the communication and control among different genes. In accordance with Wiener's definition of classical cybernetics [298], I will call this a *cybergenetic* system, and draw inspiration from general cybernetic theory throughout this thesis. Finally, note that for each of the processes in Figure 1.1, there is also a process in the opposite direction. RNA can be reverse transcribed into DNA (though this primarily happens in viruses), proteins change the sequence structure of RNA molecules in a process called splicing, and the higher-level organisation of DNA can affect protein binding and cooperativity [128, 134].

1.2 Gene expression and regulation

The abstraction generally used to refer to differences in concentrations of gene products is called *gene expression*. Genes themselves are already abstractions of the discrete units of inheritance, but many DNA, RNA, and protein sequences can be associated with a particular gene by matching sequences through the complementarity of nucleotides and

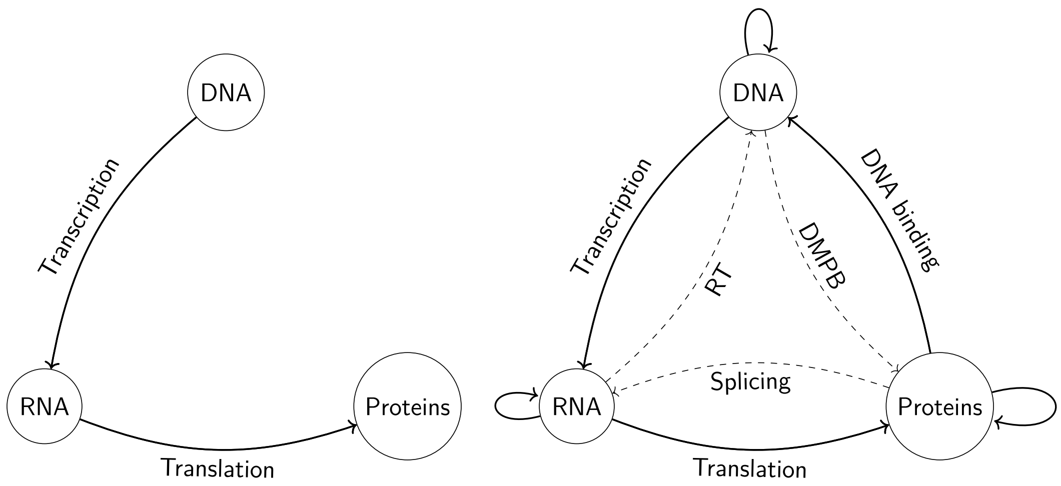


Figure 1.1: Sequence information flow in the central dogma (left), versus more general *in vivo* observed regulatory information flows (right), which include reverse transcription (RT), RNA-splicing by proteins, DNA-mediated protein binding (DMPB), and DNA-binding by proteins.

the genetic code. Each gene can thus be assigned an expression level by measuring the concentrations of the corresponding molecules, either RNA or proteins. Good evidence that it is indeed these relative concentrations of molecules that differentiate cell types comes from so-called cell atlases, which cluster cells by their gene expression profiles, and will be discussed in more detail in Section 1.3. This clustering often matches with microscopically observed cell types, and can even reflect differentiation dynamics. Regulatory networks can dynamically change the expression profiles within cells, but preserve transcriptional states across mitosis [120], leading to the stable diversity seen across the cell types in an organism.

Molecular biology of gene expression We say that a gene is expressed inside a cell if any of the gene's products, be they RNA molecules or proteins, are present or being produced inside the cell. For a gene to be expressed, it first needs to be transcribed. The process of transcription is divided into three steps: initiation, elongation, and termination. During initiation, an RNA polymerase (RNAP) molecule must bind to the DNA at the correct location near the gene, which is facilitated by the presence of so-called promoter regions. A promoter is a region of DNA that allows RNAP to bind stably to the DNA and start transcribing a particular gene. How RNAP finds the promoter region in the first place is an issue of active research and debate. It is generally assumed that most biochemical processes inside the cell are stochastic in nature, and that the rate and location at which proteins bind to DNA are driven by thermodynamic diffusion effects [95], either 1-dimensional diffusion along the DNA strand, or 3-dimensional diffusion through the nucleus. However, recent studies have suggested that liquid-liquid phase separation can occur at specific locations along the DNA, allowing for more controlled and directed diffusion of proteins like RNAP to promoter regions [33, 54]. In both these cases, the accessibility of the promoter regions is crucial, so the regulatory degrees of freedom that determine the rate of transcription are the chemical and 3-dimensional configuration of the DNA molecule and the concentrations of molecules like RNAP. Certain proteins, called transcription factors (TFs), can alter the binding affinity of RNAP to the promoter by binding to the DNA. Some TFs do this by directly binding to the promoter sequence to affect the binding of RNAP, while other TFs bind so-called *enhancer* regions, located elsewhere in the genome. These enhancers can be cis-acting on genes in the genomic neighbourhood, or trans-acting on genes several million base pairs (Mbp) or chromosomes away. Enhancer regions can help recruit RNAP to the promoter region by physically changing the geometry of the DNA molecule. On top of these regulatory mechanisms, the DNA molecule can contain chemical epigenetic modifications, which change the accessibility of the affected regions, affecting the rate at which transcription can take place. This leads to a model of transcription in which the rate at which a certain gene is transcribed depends on the configuration of the DNA, which in turn depends on the presence and activity of many proteins, most notably RNAP and the TFs that target that particular gene.

Once RNAP has bound to the promoter region and transcription has started, the elongation phase begins, and RNAP traverses the DNA molecule in the 3' to 5' direction of the template strand, forming the complementary RNA molecule in the 5' to 3' direction. The speed at which RNAP generates the RNA, as measured in bases per unit of time, is variable, regulated, and consequently affects the rate of transcription [172].

At the end of the gene, there is a region called the terminator, which releases the generated RNA molecule from RNAP upon transcription. In eukaryotes, the released RNA molecule is called a pre-mRNA molecule, as it could still be modified by a process called splicing before it gets translated. Not only can splicing alter the rate at which a gene product becomes functional, it can also create different versions of the same genes, called splice-variants, by splicing and combining the exons in different ways. While the RNA concentration is usually what defines a gene's expression level, most genes' primary function is in the form of proteins. To translate the processed (or *mature*) mRNA into proteins, the transcripts are transported to the cytoplasm, where ribosomes bind to the RNA and synthesise the corresponding protein from amino acids, using transfer RNA (tRNA) to read out the genetic code. The rate at which the ribosomes bind to and translate the RNA is controlled mainly by the presence of proteins called initiation factors, and by regulatory sequences on the mRNA molecule, most notably the 5' untranslated region (5' UTR). Finally, the mRNA molecules are degraded at a rate that is controlled by their sequence, and the rate of translation, most notably by the trimming of their stabilising polyadenylated tail by exonucleases.

Gene regulatory networks These regulatory mechanisms give a general description of gene regulation in which gene expression is controlled by the concentration of RNA and proteins, which are themselves gene products. This leads to the notion of a *gene regulatory network* (GRN), which is a summary of these regulatory relationships among genes. In a GRN, the information at which biochemical level the regulation occurs is usually abstracted away, and one only talks about genes and their regulatory relationships. For example, a transcription factor B might bind to an enhancer region that increases RNAP recruitment to the promoter region associated with a gene A. In the GRN, this could simply be summarised as an activating relationship $B \to A$. If B is itself the target of another transcription factor C, the GRN would contain the chain $C \to B \to A$. In a GRN, it is not clear if $B \to A$ means that the protein B directly binds to the promoter, if it recruits RNAP by binding to an enhancer, or if it increases the transcription of long noncoding regulatory RNA molecules that affect A (so-called lncRNAs [282], including enhancer-derived RNA [205]). Each of these biologically different situations leads to a gene-gene relationship where the expression of B affects the expression of A. How to abstract the biochemistry into a GRN is the central question in the construction of GRNs, and there are many proposed methods—each with its own (dis)advantages—all relying on the intuition that a regulatory relationship will reveal itself in the relative concentrations of the gene products [75]. Some of the most common techniques to construct GRNs from expression data are introduced and discussed in Section 1.5. Not all regulatory pathways are active in each cell, so which relationships are found will depend on the data used to construct the GRN. When inferring a GRN on data from a single tissue, the GRN will contain both organism-wide, and tissue-specific regulatory relationships, but pathways that are specific to only one of the cell types in the tissue might be averaged out and disappear, or be cancelled by the opposite regulatory relationship in a different cell type [216, 273, 171]. A given GRN is thus only interpretable relative to the context in which it was constructed. This makes validation against known biology challenging, since many of the expert-curated databases of regulatory relationships, like the STRING [264] or Pathway Commons [219] databases, integrate multiple different sources of information, and do not specify tissue- or cell-type-specificity. In the absence of a clear ground-truth network, it is difficult to say exactly how a GRN should be judged. An alternative to using a ground-truth network is combining known biological annotations of genes with the intuition that genes with a similar annotation should be more likely to interact. Such biological annotations are summarised in gene-ontology (GO) databases [56]. Methods to compare such annotations will be discussed in more detail in Section 4.1.1.

Still, the question of what kind of biology should be captured by GRNs is unanswered. Certainly, edges in a GRN should reflect direct regulatory relationships like a transcription factor binding to its target's promoter region. But should the binding of two proteins to form a complex be reflected as a regulatory relationship in the network? Should protein function be reflected in RNA concentration? These are open questions, and different GRNs address them in different ways. Simpler to state is the requirement that the interactions in the network be direct. That is, if a particular biochemical pathway decomposes as $A \to B \to C$, then this should result in the edges $A \to B$ and $B \to C$ being present in the model, but not the edge $A \to C$. More generally, this means that the edges should contain causal information. In his work on causal inference [190], Pearl describes the different levels of causal reasoning: statistical, interventional, and counterfactual. Certainly, all models should accurately capture statistical relationships. Ideally, a network could also make interventional predictions that would allow for targeted therapies in disease, and provide control over cell and tissue differentiation. Furthermore, such predictions can be explicitly falsified by interventional experiments. Counterfactuals require not only a causal model, but also functional dependencies and detailed knowledge of unobserved confounders, so are probably out of reach for most networks that only use one data source.

Mammalian development and cell differentiation As an organism develops, a single zygote must give rise to many different kinds of cells and tissues. It does so through a combination of cell division and gene regulatory programs. Crucially, as a cell divides, its daughter cells inherit part of an epigenetic state, which includes (but is not limited to) the transcriptome, proteome, metabolome, and methylome, although cell division is not necessarily symmetric. After fertilisation, a zygote enters a cell state that sets off a series of changes in gene expression that, upon cell division, leaves the daughter cells in different states. These daughter cells will in turn develop differently due to the regulatory dynamics between the molecules they contain, and their different environments and internal states. Since both the molecules and their regulatory relationships are determined by the genome, it is in this sense that the whole developmental programme is encoded in the genome and the initial state of the zygote. The zygote can generate all other cell types, and is therefore called *totipotent*. In many animals, the zygote first forms a *blastula* (in mammals also called a *blastocyst*) of many totipotent cells, before rearranging itself into three *germ layers*—the endoderm, ectoderm, and mesoderm—in a process called *gastrulation*. These will form the inner- and outer boundaries and the inside of the organism, respectively. Cells in these three layers are no longer totipotent, but will go on to generate the many different cell types in the body, so are still called *pluripotent*. As such, each cell in a body has a *lineage*, a sequence of cell states that traces its origins, starting from the zygote. The process by which a cell traverses this lineage towards a cell type is called cell *differentiation*. In general, differentiation is a process of specialisation, as more and more cell states are excluded as future cell fates. A cell at the end of a lineage that can no longer specialise further is called *terminally differentiated*. However, there are important exceptions to this general pattern. Under the right circumstances, cells can also dedifferentiate to an ‘earlier’ state in the lineage. For example, the introduction of just four transcription factors, the Yamanaka factors, is enough to dedifferentiate common somatic cell types like fibroblasts into fully pluripotent stem cells, called induced pluripotent stem cells (iPSCs). Charting the landscape of the different cell types and states is thus a challenging task. Section 1.3 introduces the central ideas and techniques currently used to define and identify different cell identities, and anticipates one of the central contributions of this thesis: a new way to identify cell states in a population of single cells.

Embryonic and adult neurogenesis As an interesting example, and because this lineage will turn out to be relevant to the final chapter of this thesis, I want to highlight one particular lineage: embryonic and adult neurogenesis in mice.

Neuroepithelial cells derive from the ectoderm early in embryonic development and differentiate into radial glial cells (RGCs) [160]. That such cells play an important role in neurogenesis has been known since the late 19th century [208], but their precise role and functioning remained unclear for almost a century. Early images from electron microscopes suggested that immature neurons are generated and then mature as they migrate to their final destination along a cellular scaffolding of radial glial cells (RGCs) [207]. However, subsequent experiments showed that perhaps the main role of the RGCs is to serve as the precursor cells from which most neurons and astrocytes in the central nervous system derive [158, 179]. RGCs localise in a region of the brain known as the ventricular zone (VZ), but they divide asymmetrically to produce neuronal intermediate progenitor cells (nIPCs) which localise in the subventricular zone (SVZ) [181]. These two zones then form the main neurogenic regions during brain development. In mice, the production of neurons from these two regions and cell types then starts as early as eight days after conception (E8), but in most regions of the brain around E10, and peaks around E14 [97]. However, some neurons, astrocytes, and oligodendrocytes might already be produced by neuroepithelial and early radial glial cells [139]. Due to this multipotency, RGCs are also referred to as neural stem cells (NSCs). As RGCs mature and have produced most neurons and neural precursors, they mostly transform into astrocytes [139]. Throughout development, different regions of the brain generate different kinds of neurons. The two main classes of mature neurons are excitatory neurons, which mostly derive from the developing pallium in the ventricular zone, and inhibitory GABAergic neurons, which mostly derive from a transitory structure in the SVZ known as the ganglionic eminence [77, 240]. A subregion of the ganglionic eminence, known as the lateral ganglionic eminence (LGE), is the source of a particular lineage that produces medium spiny neurons and a population of interneurons that will migrate to the olfactory bulb. I found evidence of cells from this lineage in a population of developing neurons which is presented in Section 5.3.5 of this thesis.

It was commonly thought that neurogenesis was restricted to the embryonic stages of development, but evidence of postnatal neurogenesis was found in rodents in the 1960s [246, 10]. It was found that cells commonly thought to be normal astrocytes were actually astrocyte-like RGCs (also called radial glia-like cells) that still retained their neurogenic ability, in particular in the SVZ and a region in the hippocampus known as the subgranular zone (SGZ) [139]. These two regions are known as the adult neurogenic niches, as they provide the microenvironment in which the radial glia-like cells retained their NSC identity [30]. Evidence for neurogenic niche activity in a population of late-developmental astrocytes is presented in Section 5.3.4 of this thesis.

1.3 Cell identity: states and types

> Ignoring causation invites disaster.
>
> Dogen Zenji [265]

Cell types and atlases The dynamic nature of cell differentiation highlights the difficulty in defining cell identities as distinct and stable cellular phenotypes. It is generally assumed that genetic programmes drive development and differentiation, that the programmes function by generating gene products at different rates, and therefore that the concentrations of different gene products within the cells characterise the different cell identities present in the population. Furthermore, it is then assumed that cells of the same type have a similar expression profile *on average*, and can thus be found by clustering cells by their transcriptional profile. Genes that are most predictive of cluster identity are thought to be characteristic of that cell identity, so the *marker genes* for a particular identity are the genes that are expressed at different (typically higher) levels in that cluster relative to all other clusters. Such genes are called *differentially expressed* (DE). Typically, each cluster's DE genes are then compared with a reference list of marker genes for cell identities to assign biological meaning to the clusters (in what is called the marker-based approach to cell identity annotation). Alternatively, an already annotated population of a particular cell type can be used to compare a cell population with the pre-annotated cells from the reference population (the correlation-based, or supervised approach to cell identity annotation). Popular software packages that automate this annotation process are *e.g.* `SingleR` [13], `scmap` [135], `CellAssign` [315] and `SCINA` [318]. Note that using the same data to define the clusters and to find the differentially expressed genes artificially inflates the number of significant marker genes in a process known as *double-dipping* [88]. This is an often overlooked problem of cluster-based approaches and will be further addressed in Section 5.2.4.

The biological annotation of a cluster of cells is usually referred to as a cell type. Creating a taxonomy of cell types thus requires discretising expression space, but this is not *a priori* biologically justified. For example, two regions in expression space could be separated only by the expression of genes that are not relevant to a cell's type (like cell-cycle related, or housekeeping genes). Before genome-wide expression data was available, cell types were hypothesised to be discrete attractors of the regulatory dynamics, described by a 'very constrained pattern of gene expression' [131], but modern transcriptomic studies refute the presence of such a clear discrete landscape [273], and especially during development cells occupy a continuum of states along developmental *pseudotime* [274]. While one can use knowledge of developmental trajectories to impose a direction of pseudotime on the cell population, inferring dynamics from a static snapshot of a cell population has inherent limitations [295]. Consequently, modern cell atlases are not algorithmically constructed but require the careful yet subjective analysis of experts to identify cell types based on gene expression, cellular context, morphology, cross-species homology, perturbation experiments, lineage tracing, etc. [171]. Such (partial) atlases have for instance been constructed for the mouse *Mus musculus* [57], the jellyfish *Clytia hemisphaerica* [51] and *Nematostella vectensis* [235], the worm *Caenorhabditis elegans* [47], and different consortia are working on the primate *Homo sapiens* [58, 213]. However, crucially, all these atlases are fundamentally based on a discretisation of gene expression space, and annotate all cells in a given region as the same cell type.

Identifying cell states Throughout, and even after development when they are terminally differentiated, cells can show dynamic behaviours that preserve their cell type but change their function, morphology, or location. These changes are typically transitory and not large enough to warrant being called a different cell type, so are referred to as *cell states*. There is no clear difference between cell state and cell type, but a cell's state is generally considered a more fine-grained description than a cell's type, based on the transitory processes happening inside the cell [171]. Different cells from a single cell type can be in different states. For example, T-cells can be in a resting or an activated state [314]. However, just as T-cells can be in different states, there are also different stable (sub)types: Helper/CD4+ T cells, cytotoxic/CD8+ T cells, and regulatory T cells. The notions of cell state and type are thus not clearly separable. Because of this, some notions of cell state transcend cell types and can be spread throughout expression space. For example, different cell types in a given tissue can be in a cycling state at the same time. This notion of cell state depends more on the activity of a particular set of genes or pathway than on the average gene expression. Similar regulatory processes could still preferentially lead to similar transcriptional profiles, so cell states might still localise in expression space, but their definition cannot be based on clustering in expression space.

Identifying cell states in a population of cells is made easier by knowing in advance the cell states of interest, and which genes to base the annotation on, so is most commonly done for cell-cycle related states and, by extension, cancerous states. The authors of [115] annotate cells to states along the cell-cycle, and find that they require the expression of just five genes to do so accurately. Furthermore, including additional genes did not improve the accuracy. In the absence of knowledge of the relevant cell states and genes, dimensionality reduction techniques can also be used to find substructure in the data. In [201] and [18], the authors use nonnegative matrix factorisation (NMF) to find co-expressed gene modules that define delocalised (in gene expression space) cell states, while the authors of [188] use principal component analysis (PCA) to find state-associated gene modules. A gene-agnostic approach is taken by the authors of [175] who define cancerous states by clustering only on genes that were consistently DE across tumours, and defining gene expression relative to a set of control genes. Another popular framework to cluster cells by module activity, based on coexpression modules of TF-target pairs, is SCENIC [7]. All these approaches are based on identifying interesting modules of genes that show non-random expression in the population to identify the cell states. In this thesis, and in particular in Chapter 5, I will use the notion of a model-free interaction to automatically detect cell states based on higher-order conditional dependencies, that can delocalise in gene expression space and allow each cell to be in multiple states. Furthermore, these higher-order interactions also contain mechanistic and regulatory information, can define states by both the absence and the presence of gene products, and induce a hierarchical structure on the cell states.

Cybergenetic semantics Defining cell states through their internal regulation is a ‘systems’, or cybernetic view on cell state, where relationships and regulation are deemed at least as fundamental as the instantaneous value of any variable. By adopting the cyclic view of regulation from Figure 1.1, the cell is represented as a unitary homeostatic system—an autopoietic machine. These are defined in [161]:

> *An autopoietic machine is a machine organized (defined as a unity) as a network of processes of production (transformation and destruction) of components that produces the components which: (i) through their interactions and transformations continuously regenerate and realize the network of processes (relations) that produced them; and (ii) constitute it (the machine) as a concrete unity in the space in which they (the components) exist by specifying the topological domain of its realization as such a network.*

Genes form the components that are both the cause and the effect of the processes inside the cell—and as such constitute a strange loop [111]. I will define cellular states through this cybernetic lens on cells, using only regulatory but acausal interactions among the genes. I focus on acausal interactions because causal inference on gene expression data is too hard, but it is hard *precisely because* causality is a slippery concept in autopoietic systems. Causal inference on gene expression might not be a well-defined question. The causal power a butterfly has in triggering a hurricane is often mentioned to emphasise how complex causal networks are, but I would argue that it illustrates very well that causality is not just complex, but *nonsensical* in sufficiently complex systems—that causal questions in complex systems do not have interesting or actionable answers. In fact, that a departure from causality can bring you closer to a cybernetic/autopoietic view on biological systems was already anticipated in [161]. In its introduction, Maturana reflects on his use of causal language:

> *I made a concession which I have always regretted. I submitted to the pressure of my friends and talked about causal relations when speaking about the circular organization of living systems. To do this was both inadequate and misleading. It was inadequate because the notion of causality is a notion that pertains to the domain of descriptions, and as such it is relevant only in the metadomain in which the observer makes his commentaries and cannot be deemed to be operative in the phenomenal domain, the object of the description. It was misleading because it obscured the actual appreciation of the sufficiency of the notion of property as defined by the distinctive operation performed by the observer when specifying a unity, for the description of the phenomenal domains generated by the specified unities. It was misleading because it obscured the understanding of the dependency of the identity of the unity on the distinctive operation that specified it. It was misleading because it obscured both the understanding of the phenomenal domains as determined by the properties of the unities that generate them, and the non-intersection of the phenomenal domains generated by the operation of a composite unity as a simple unity in a medium and by the operation of its components as components.*

Throughout this thesis, I will be oscillating between this *domain of descriptions* and *metadomain of commentaries*, so will inevitably be using causal language. In fact, a significant portion of this thesis will be dedicated to the identification of causal relationships, as they will, paradoxically, be essential in the estimation of the non-causal networks. Still, I hope to present a view on regulation and cell identity that transcends causal language, and develop a cybergenetic semantics of cell state.

1.4 Measuring gene expression

> If you try and take a cat apart to see how it works, the first thing you have on your hands is a non-working cat.
>
> Douglas Adams [5]

A gene's expression can be quantified in terms of the concentration of its products, so to measure gene expression, one has to measure the concentrations of some or all of the relevant gene products. Most gene regulation affects the rate of transcription, which is most directly reflected by the concentration of RNA molecules (though steady-state mRNA abundance is not always directly correlated to the rate of transcription [89]), but regulation is often implemented by transcription factors, which are proteins. Across species and experiments, it has been noted that the correlation between the concentration of mRNA and the corresponding protein typically is low [26, 285], an issue generally attributed to the wide range of protein degradation rates [156]. Measuring gene expression as RNA or protein concentration can thus lead to different descriptions of the same cell. A complete description of gene expression thus requires simultaneous measurements of the whole transcriptome (all transcripts present in the cell), and the proteome (all proteins present in the cell). However, single-cell measurements of the proteome generally require mass-spectrometry analysis, and are thus more difficult than measuring the transcriptome [284], which can be done using the massive parallel sequencing techniques developed for DNA. As a result, single-cell measurements of the proteome are currently limited to a few thousand genes in hundreds of cells [168], or around a thousand genes in thousands of cells [232]. In contrast, the (polyadenylated) transcriptome can be measured in a genome-wide way in over a million cells in parallel [4]. In this thesis, I decided to focus on the transcriptional profiling of cells while ignoring protein abundances. This can in part be justified by the focus on regulation, the effects of which should be most directly measurable in RNA concentrations, but it is also a practical decision as transcriptomic data is more readily available.

To measure the transcript abundance across the genome, one first needs to isolate the RNA from the cells in the sample. This RNA is then reverse transcribed into the complementary cDNA molecule, at which point it can be PCR-amplified and sequenced using DNA sequencing technology. The found sequences, called *reads*, can then be compared to a reference genome to assign each read to a gene. This is not trivial, because the amplified transcripts are first cut into small fragments—typically of a few hundred base pairs—that can be more easily sequenced, which complicates the subsequent genome alignment. Finally, a gene's expression can then be quantified as the relative abundance of the reads that mapped to that gene. The collection of methods that use a protocol like this is referred to as RNA-seq [293], and different RNA-seq experiments differ in the way they isolate the RNA from a sample, how they prepare the cDNA (called library preparation), and how the library is sequenced. In this thesis, I will focus on gene expression measurements from single cells, using the droplet-based library preparation of the *10X Chromium* system to process hundreds of thousands of cells in parallel [2]. By tagging each transcript with a unique molecular identifier (UMI), and sequencing from a primer at the 3' end, the Chromium protocol eliminates the bias in the read counts as a result of gene length or PCR amplification. The Chromium system isolates all RNA molecules with a polyadenylated tail, which includes all protein-coding genes and many non-coding regulatory RNAs, and attaches TruSeq adapter sequences to the captured transcripts to prepare them for Illumina sequencing. This provides genome-wide gene expression profiles for all cells in the sample. There is a tradeoff in this protocol where increasing the total number of cells results in a more shallow sequencing of the library for a fixed number of total reads. This means that especially in data sets with many cells, the data is often artificially sparse, and many genes get incorrectly annotated as not expressed—a phenomenon known as *dropout* [133]. How to interpret the gene expression profiles in light of this inflated proportion of zeros is an area of active debate, and opinions range from dropout being something that needs to be corrected for [210, 39, 281] to it being a signal as useful as the actual read count [204, 11, 37]. Besides dropout, there are other sources of technical and biological noise that have to be taken into consideration, which will be described in more detail in Sections 2.2.5 and 4.2.6. Only once the gene expression data has passed all these quality control (QC) steps, is it ready to be analysed. Section 1.5 introduces some of the most common techniques to construct a GRN based on expression data.

1.5 Models of pairwise interaction

Given a set of expression data across samples, be they organisms, tissues, or cells, reconstructing a regulatory network involves capturing the statistical dependencies among the genes. As the most common type of genome-wide expression data is observational data at a single time point, I will focus the discussion here on inferring GRNs from such data, not addressing dynamical inference on time-series data. There are many ways to describe and define a statistical dependency, and I will describe some of the most common techniques here. What all of these methods have in common is that they describe *pairwise* relationships between genes. That is, they assign a regulatory relationship to a pair of two genes. This makes them all suitable to a description in terms of a mathematical graph:

Definition 1 (Graph)**.** *A graph $\mathcal{G}$ is a tuple (V, E), where V is a set of vertices, and $E \subseteq \{(v, w) \mid v, w \in V\}$ is a set of pairs of vertices, and each such pair is called an edge. If E consists of unordered pairs, we call $\mathcal{G}$ an unoriented graph. If E consists of ordered pairs, we call $\mathcal{G}$ an oriented graph.*

Definition 2 (Weighted graph)**.** *A weighted graph $\mathcal{G}$ is a tuple (V, E, w), where (V, E) forms a graph, and w is a function $w : E \to \mathbb{R}$ that assigns a weight to each edge.*

Consider the weighted graph $\mathcal{G} = (V, E, w)$ with $E = \{(g_i, g_j) \mid (g_i, g_j) \in V \times V\}$, called the complete graph on V. When V is a set of genes, and E describes the relationships between them, I will refer to $\mathcal{G}$ as a gene regulatory network. The different ways to construct a GRN correspond to different choices and interpretations for the weight function w. I will describe four different classes of GRN construction methods here: coexpression networks, regression networks, Bayesian models, and Ising models, and highlight some relationships between them.

1.5.1 Coexpression networks

Coexpression networks are non-parametric GRN inference methods that assign a weight w_{ij} to each edge (g_i, g_j), based on some coexpression pattern of the two genes. There are different ways to quantify coexpression, four of which are discussed below.

Correlation The most straightforward way to create a genetic network based on expression is by pairwise Pearson correlation. Computing the correlation for each pair of genes results in a fully connected, undirected graph of statistical association. Such networks are sometimes simply called coexpression networks. To use quantities based on the network topology, one can introduce a threshold on the edges to end up with a network of just strongly correlated genes, though simply thresholding on correlation strength is known to induce spurious structure [46]. A commonly used framework is [316]. Since correlations are only sensitive to linear relationships, nonlinear pairwise interactions can result in false negatives. False positives can appear whenever a third gene interacts with two genes that do not interact among each other, since correlations do not disentangle direct from indirect effects.

Partial Correlation To calculate partial correlations, $g_i, g_j \in G$ are first regressed against $G \setminus \{g_i, g_j\}$, and the correlation is calculated on the residuals. For jointly Gaussian data, this partial correlation exactly captures the conditional dependence structure. Partial correlations can thus disentangle direct and indirect associations, at the cost of assuming Gaussianity, and still assume linearity. In practice, one often does not have to condition on all other genes, and a correlation between residuals after a regression on n genes is referred to as an nth-order partial correlation [64]. If the genes' expression levels are described by a multivariate normal distribution, then the inverse covariance matrix (if it is defined) contains precisely the partial correlations, which is why partial correlation networks are sometimes also called gaussian graphical models [140].

Distance Correlation Alternatively, to get rid of the linearity constraint, distance correlations can be used, which are sensitive to nonlinear relationships [185, 263]. The authors of [262] combined these ideas and used a partial distance correlation for GRN inference.

Mutual information Another attractive, nonlinear alternative to correlations is to use quantities based on information theory, like mutual information (MI). MI is sensitive to nonlinear relationships, and can be conditioned on any number of genes, but is still a symmetric measure of association so results in an undirected graph. Mutual information is arguably the canonical definition of dependence between two variables, as it is the KL-divergence between the joint distribution and the product of the marginals. Popular GRN inference techniques based on MI are ARACNE [159], and context likelihood of inference (CLR) [79]. MI has a natural extension to beyond-pairwise dependencies that will be introduced, and play an important role, in Section 3.3.1. Higher-order mutual information has been used in combination with CLR to win the DREAM2 challenge of inferring GRNs from expression data [292].

1.5.2 Regression

In coexpression networks, the edges are all undirected, and there is no notion of prediction. To introduce a direction to associations, each gene g_i is regressed against all other genes g_j, and coefficients β_{ij} that have an inherent direction and weight $g_j \xrightarrow{\beta_{ij}} g_i$ are obtained. These regression coefficients predict the expression of a gene from the other genes' expression. The DREAM4 challenge was won by an algorithm that used such regression coefficients together with a tree-based ensemble method to choose the most important coefficients [119]. Due to its simplicity and intepretability, regression is used in many different contexts. In [288], the authors construct networks of gene-gene associations by regressing phenotypic profiles across experimental conditions. While this is a fundamentally different kind of data, the resulting network is similar and complementary to expression-based GRNs.

1.5.3 Bayesian models and causal DAGs

The joint probability of an expression profile over all n genes $p(G) = p(g_1, \ldots, g_n)$ can be factorised using the chain rule for probabilities:

$$p(G) = p(g_1 \mid G \setminus g_1)\ p(g_2 \mid G \setminus \{g_1, g_2\})\ p(g_3 \mid G \setminus \{g_1, g_2, g_3\}) \ldots p(g_n) \tag{1.1}$$

This is always true, but can be simplified further when not all genes directly affect each other. Note that if two genes g_a and g_b are conditionally independent given g_c, then

$$g_a \perp\!\!\!\perp g_b \mid g_c \implies p(g_a \mid g_b, g_c) = (g_a \mid g_c) \tag{1.2}$$

Such conditional independencies can reduce the conditioning sets in each of the terms in Equation (1.1). If there exists an assignment of a set of genes $\mathrm{Pa}(g_i)$ to each gene g_i—called its parents—such that

$$p(G) = \prod_{i=1}^{n} p\left(g_i \mid \mathrm{Pa}(g_i)\right) \tag{1.3}$$

then the distribution in Equation (1.3) is called a Bayesian network and has an associated oriented graph $\mathcal{G} = (G, E)$ where $E = \{(a, g_i) \mid g_i \in G,\ a \in \mathrm{Pa}(g_i)\}$, that is, each gene has incoming arrows from all its parent genes.

While each joint probability distribution can be assigned a decomposition graph like this, not all graphs correspond to a valid decomposition of a joint distribution. In particular, a graph with cycles does not allow for a definition of parents that makes Equation (1.3) a Bayesian network, so Bayesian networks correspond to *directed acyclic graphs* (DAGs). Furthermore, not every joint probability can be assigned a unique DAG. For example, $A \to B \to C$ and $A \leftarrow B \leftarrow C$ are two different DAGs that correspond to the same conditional independencies. A given conditional independency structure can thus only be assigned a *Markov equivalence class* of DAGs that encode the same dependency structures. Once the structure of the Bayesian model is determined, the conditional probabilities can be estimated from data, but the greatest challenge in practice is identifying all conditional independencies in the data. The procedure of finding the graph of conditional dependencies goes by the name of causal discovery, since such a graph can, up to Markov equivalence, reveal the causal dependencies in the data. If there are few variables, all possible graphs can be efficiently ranked by their marginal likelihood. This is called the score-based approach to causal discovery. However, the number of possible DAGs scales superexponentially with the number of nodes [218, 245], so this quickly becomes intractable. Alternatively, conditional independence tests can be used to remove edges from the complete graph, as is done in the Peter-Clark (PC) algorithm which will be described in Section 4.2.4. The PC-algorithm is an example of a constraint-based approach. Both approaches can determine the Markov equivalence class of graphs that are consistent with the observed joint probability. Bayesian networks can reveal many of the dependencies in the data in a non-parametric way, while assigning a causal direction (up to Markov equivalence) to each of the associations. However, this does come at the cost of assuming that the true causal structure underlying the data is indeed acyclic. While this assumption is necessary for almost all reductionistic views of causality, it is wholly unrealistic for GRNs, where feedback loops are common, and biologically necessary.

Traditionally, the restriction of causal reasoning to acyclic systems is seen as a technical limitation—we simply lack the mathematical tools to describe the causal effect of an intervention, or to discuss counterfactuals, in situations with cycles or symmetric interactions. While there have been advances in the field of causal discovery with cyclic graphs (see *e.g.* [83]), I argue that this is not just a technical challenge, but rather indicative of the limits of causal reasoning itself. For example, intervening in a negative feedback loop negates the intervention itself. Describing only the short-term behaviour of the variables in the loop means 'cutting' the loop and making it a chain, thus not accurately describing the cyclic system. Describing the long-term effects of the intervention involves negating the intervention, so describes both the effects of your intervention and its negation, which makes it meaningless. It is in this way that causal questions in a cyclic system can lead to a kind of liar-paradox—a conundrum that has heralded breakdowns of many other formalisms as well. Furthermore, causal reasoning is fundamentally based on the principle of *ceteris paribus*—all other things being equal—which is a state that a sufficiently complex system will never reach. Still, many complex systems have descriptions in terms of causal explanations of behaviour. For example, the introduction of a predator can causally break up a school of fish. However, the causal variables in such descriptions are part of an emergent, coarse-grained description of the system, in which the symmetric or cyclic microscopic interactions are abstracted away. Formal descriptions of such higher-order causal variables have recently been developed in *e.g.* [8, 110]. These approaches emphasise the role of emergent descriptions and variables necessary to apply causal thinking to complex systems. Alternatively, one can abandon directionality and causality, as is done in many areas of physics. One example of this is the final class of statistical models introduced here: the Ising model and its generalisations.

1.5.4 Ising- and other spin models

> Which way does the arrow point?
> This uncertainty is healthy.
>
> Merlin Sheldrake [237]

1.5.4.a Physical interpretation of the Ising model

The Ising model was originally introduced to explain the behaviour of crystalline magnets, and describes the behaviour of the magnetic moments of atoms. It does so by assigning an energy E to each configuration s of the crystal, composed of N individual atoms with magnetic moments s_i:

$$E(s) = \sum_{i=0}^{N} h_i s_i + \sum_{i=0}^{N} \sum_{j=0}^{i} J_{ij} s_i s_j \tag{1.4}$$

which describes the energy of classical atoms in an external magnetic field h_i, with a pairwise interaction term J_{ij}. At equilibrium, each configuration s occurs with a probability given by its Boltzmann weight:

$$p(s) = \mathcal{Z}^{-1} \exp(-E(s)) \tag{1.5}$$

where $\mathcal{Z}$ is the normalising factor $\sum_t \exp(-E(t))$. Usually, one assumes that the variables s_i are binary, the external magnetic field is homogeneous ($h_i = h$), and that the atoms form a 1- or 2-dimensional lattice where only nearest neighbours interact, setting $J_{ij} = J$ whenever s_i and s_j are nearest neighbours on the lattice, and $J_{ij} = 0$ otherwise. This is the system that has historically been referred to as the Ising model. Deriving properties of this distribution for a given field strength h and nearest-neighbour coupling J is known as the *forward Ising problem*, and has led to great insight into the behaviour of atoms.

If the fields and interactions are allowed to be inhomogeneous, and the s_i are real-valued, then Equation (1.5) with the energy function from Equation (1.4) just describes a Gaussian distribution with a covariance matrix completely specified by h_i and J_{ij}. In fact, the couplings are just the entries in the inverse covariance matrix (when it is well-defined). This gives an interpretation of the Ising model as a Gaussian graphical model, and thus in terms of partial correlations. Due to the inhomogeneity, such models are also referred to as spin, or spin-glass models.

1.5.4.b The Ising model as maximum entropy inference

That the Ising model shares some structure with statistical models is no coincidence. In 1957, E.T. Jaynes showed that statistical mechanics, and in particular the Ising model, can be seen as the solution to an inference problem [127]. Consider a system of N variables $X = \{X_1, X_2, ..., X_N\}$, where X takes values in a discrete state space $\mathcal{X}$. After observing independent and identically distributed (*i.i.d.*) samples of the system, one might want to write down a model for X. Without prior knowledge about the dynamical laws, the most general solution is to assign each of the possible configurations a probability, *i.e.* describe $p(X = x) \quad \forall x \in \mathcal{X}$. To be maximally noncommittal with respect to unknown properties of p, the entropy $H(p) = \sum_{x \in \mathcal{X}} p(x) \log p(x)$ should be maximised, subject to some constraints. What kind of constraints should be enforced? The function p should certainly be a well-defined distribution, so probabilities should sum to one. To enforce this, write $p(X = a) = p(a)$, introduce a Lagrange multiplier λ_0, and set the following quantity to zero:

$$\frac{\partial}{\partial p(x)} \left(-\sum_{y \in \mathcal{X}} p(y) \log p(y) + \lambda_0 \left(\sum_{y \in \mathcal{X}} p(y) - 1 \right) \right) = 0 \tag{1.6}$$

which forces

$$p(x) = \exp(\lambda_0 - 1) \implies \lambda_0 = \log(|\mathcal{X}|) + 1 \tag{1.7}$$

such that for all $x \in \mathcal{X}$

$$p(x) = \frac{1}{|\mathcal{X}|} \tag{1.8}$$

This is just the uniform distribution over all possible states, and reproduces the intuition of the principle of insufficient reason [65]. If the expectation value μ_j of each variable

X_j should also be reproduced, as well as the expectation values σ_{ij} of the products X_iX_j (*i.e.* the first two moments of X), additional Lagrange multipliers are introduced:

$$\frac{\partial}{\partial p(x)}\left(-\sum_{y\in\mathcal{X}} p(y)\log p(y) + \lambda_0\left(\sum_{y\in\mathcal{X}} p(y) - 1\right) + \sum_{i=1}^{N}\lambda_i\left(\sum_{y\in\mathcal{X}} p(y)y_i - \mu_i\right)\right.$$
$$\left. + \sum_{i,j=1}^{N}\lambda_{ij}\left(\sum_{y\in\mathcal{X}} p(y)y_iy_j - \sigma_{ij}\right)\right) = 0 \quad (1.9)$$

where y_i is the value of variable X_i in state $X = y \in \mathcal{X}$. This has as a general solution:

$$p(x) = \exp\left(\lambda_0 + \sum_{i}^{N}\lambda_i x_i + \sum_{i,j}^{N}\lambda_{ij}x_ix_j - 1\right) \quad (1.10)$$

When the X_i are binary variables, finding the λ's that satisfy each of the constraints is called the *inverse Ising problem*, as the distribution exactly describes the equilibrium dynamics of a glass-like classical Ising model with pairwise interactions λ_{ij} and varying magnetic field λ_i. As mentioned before, when the X_i are continuous, then this is easily solved by the inverse covariance matrix, which has been applied in the context of gene regulatory networks [148, 152]. However, discrete X_i are not normally distributed, so the inverse covariance matrix does not encode the couplings.

Solving this inverse (pairwise) Ising problem has received a lot of attention, a good review of which is [178]. While the likelihood function of the Ising model is convex in the coupling parameters, evaluating the gradient involves a number of terms exponential in the number of variables, so is intractable in practice. One promising technique involves the use of a class of neural network called restricted Boltzmann machines, which will be introduced and evaluated in Chapter 2.

1.6 Higher order interactions and synergy

> You cannot answer a question that you cannot ask, and you cannot ask a question that you have no words for.
>
> Judea Pearl [189]

All definitions of genetic interactions introduced so far have one thing in common: they describe pairwise interactions only. That is because they are all based on the notion of a graph of associations, where each edge connects precisely two genes. This constraint can be relaxed, leading to the notion of a hypergraph:

Definition 3 (Hypergraph)**.** *A hypergraph $\mathcal{G}$ is a tuple (V, E) where V is a set of vertices, and $E \subseteq \mathcal{P}(V)$, where $\mathcal{P}(V)$ is the powerset of V.*

A hypergraph can contain edges between any number of nodes. For example, E can contain the triplet edge (a, b, c) while not containing the pairwise edges (a, b), (a, c), and (b, c). If edges still encode dependencies in the data, then these hyperedges encode *higher-order* dependencies. Higher-order interactions might seem exotic since they can no longer be represented by a graph, but are important or even necessary components of some systems. As a canonical example, consider the XOR logic gate. It has the following truth table for two input variables A and B, and output C:

A	B	C
0	0	0
0	1	1
1	0	1
1	1	0

This is one of the six fundamental non-trivial logic gates, and it has the property that each pair of variables is completely independent, even though the logic gate specifies a particular dependency of the output on the input. All of its structure is encoded as a third-order dependency among the unordered triplet (A, B, C). The symmetry of this interaction is reflected in the logic: the truth table describes an XOR gate for any partition of the three variables into two inputs and an output. A higher-order dependency like this is called *synergistic*, and the XOR gate (and its negation XNOR) is purely synergistic. Indeed, an Ising model on three variables $s = (s_1, s_2, s_3) \in \{-1, 1\}^3$ with only a third-order interaction $E(s) = Js_1s_2s_3$ has the property that:

$$\log \frac{p(s_3 = 1 \mid s_1, s_2)}{p(s_3 = -1 \mid s_1, s_2)} = \log \frac{\exp(Js_1s_2)}{\exp(-Js_1s_2)} = 2Js_1s_2 \tag{1.11}$$

which indeed describes the behaviour of an XOR gate. Such Ising models with higher-order interactions arise naturally as a generalisation of Equation (1.10), when moments beyond the second moment are constrained in the entropy maximisation. Demanding that the inferred distribution reproduces up to n moments leads to a generalised Ising model with interactions up to nth order.

The role such higher-order interactions play in biological networks is not yet clear. This is in part because they are harder to estimate (either mathematically, or in terms of the required statistical power), but also because pairwise descriptions of biological systems often work surprisingly well. Investigations into this phenomenon have shown that there are regimes in terms of the strength and density of couplings in which this pairwise sufficiency tends to hold [163, 271]. On the other hand, there is evidence that it is precisely the presence of these higher-order interactions that is responsible for some of the rich dynamics [21] or bistability [244] in biological networks, and synthetic lethality experiments in yeast have shown that trigenic interactions form a larger network than pairwise interactions [142]. Some gene regulation is inherently combinatorial and involves simple logical operations. Examples include the combinatorial nature of BMP signalling [12, 137] or the pattern formation in *eve* expression in developing *Drosophila* [14]. Reconstruction of GRNs using Boolean networks explicitly involves logical dependencies, and thus implicitly imposes higher-order dependencies. These networks have been popular since the 1960s [257], but are still in use today [305, 234, 110]. In the field of information theory, higher-order quantities have been in used in many contexts, and were already flagged as relevant for GRN reconstruction in [159, 176]. Furthermore, the synergistic information among variables can identify macroscopic causal quantities and descriptions [283]. It should be emphasised that I only consider those statistical dependencies that cannot be decomposed into pairwise dependencies to be higher-order. There have been claims of higher-order interactions in single-cell gene expression data before [93], but there it refers to a change in coexpression over (pseudo-)time, which could still be decomposable into pairwise interactions[2], and is not a higher-order interaction in the sense of this thesis.

It is unclear if a higher-order dependency truly reflects a microscopic logic-like interaction, or if it is an emergent dependency induced by coarse-graining or a specific choice of variables. In [222], the authors emphasise the difference between higher-order mechanisms, which are beyond-pairwise terms in the data-generating process, and higher-order behaviour, which are the emergent beyond-pairwise dependencies in a description of the data. Systems with higher-order behaviour need not have higher-order mechanisms. As an illustrative example, the authors of [222] discuss a frustrated spin system composed of three variables that all couple in pairs, but negatively. While all interactions are pairwise, the system shows a higher-order dependency (as measured by the total correlation across the three variables), while containing only pairwise mechanisms. At the same time, higher-order mechanism and behaviour are related, since especially in biology, each interaction could be decomposed further, taking you from the realm of biology, through chemistry, to fundamental physics. Any higher-order dependencies are thus a reflection of the particular way you choose to decompose your system. If you can justify a particular decomposition, then the higher-order dependencies among these variables inherit the justification. Throughout this thesis, I will describe cells in terms of gene expression. While it is clear that genes are not the only relevant variables that determine cellular dynamics (alternatives range from metabolic, epigenetic, or even bioelectric causal influences), they form a natural decomposition in which to describe cellular dynamics. Genes are a natural way to *carve Nature at its joints* [194].

However, if higher-order interactions do not directly correspond to a known biological mechanism, how are they to be validated and interpreted? This is discussed in more detail in Section 4.1.1, but key is that the biological knowledge extracted from the interactions should be meaningful, and reproducible. Their interpretation and validation should thus depend on the analysis that the interactions are used in. It will turn out that the higher-order interactions will be most useful for cell state identification, in which case validation amounts to verifying that the found cell states align with observed cell identities in the literature, but also that they are present throughout biological replicates. Finally, it should be emphasised that the interactions studied in this thesis are fundamentally undirected, which gives them no meaningful causal interpretation. While they will sometimes be compared with the directed acyclic graph that aims to capture the conditional dependencies, the two are fundamentally different quantities and should not be conflated.

[2] Consider, for example, the negative correlation between two negatively coupled genes A and B as a third gene C gets expressed more over time. If C is sufficiently coupled with positive pairwise interactions to both A and B, the correlation between A and B could become positive.

1.7 Aim and outline of this thesis

> Indeed the power and majesty of the nature of the universe at every turn lacks credence if one's mind embraces parts of it only and not the whole.
>
> Plinius the Elder [206]

Aim In this thesis, I explored the presence, structure, and role of higher-order dependencies in gene expression. The many contexts in which they have been studied have focused on specific examples, in terms of logical dependencies or synthetic lethality, but here I took a maximum entropy, purely data-driven approach. I searched for higher-order interactions in different cell types and quantified their structure, and relationship to biology, with the aim of answering the question

> *Does single-cell gene expression data contain higher-order interactions? If so, how do the interactions reflect biology?*

The first of these two questions can be straightforwardly answered by estimating the interactions on gene expression data and deciding if they are significantly nonzero. The second question—how to relate the nonzero interactions to biology—is conceptually more difficult. Fundamentally, interactions describe a conditional dependency in the expression of different genes. The most obvious way in which genes would influence each other's expression levels is by a direct regulatory mechanism. For example, one would expect a dependency between the expression level of a transcription factor and its target gene. The genes coding for two proteins that mainly function as a dimer might also be reasonably expected to show a dependency in their expression levels. Based on this intuition, I started by asking the question:

> *Do higher-order interactions correspond to complex gene regulatory mechanisms?*

To answer this question, I validated the predicted interactions against databases of known genetic interactions and annotations. However, there are other situations in which one would expect to find a dependency in gene expression. Cellular dynamics are generally attributed to genetic programmes. Taking a more agnostic approach to the biological mechanism underlying the programme, one could hypothesise that the dependency among the genes is the result of a certain regulatory programme being active in at least some of the cells. This programme could be triggered or mediated by a gene that does not directly regulate or bind to the other genes, but does influence the expression through some unobserved confounders or other biological mechanisms (like epigenetic or metabolic molecules). This more agnostic approach leads to the hypothesis that the dependencies, and by extension the interactions, correspond to cell identities marked by a combinatorial gene expression pattern, and thus motivates the second central question of this thesis:

> *Do higher-order interactions reveal cell states?*

To answer this question, I associated characteristic cell states to higher-order interactions, and compared these states with known biological cell states and types. This turned out to be the more fruitful path and led to the main results of this thesis.

Outline In practice, answering these questions required two steps: I first needed to estimate the interactions, and then validate them against known biology. To estimate the interactions, two different approaches were explored—in **Chapter** 2 a machine learning approach based on restricted Boltzmann machines (RBMs) was used, and in **Chapters 3, 4, and 5** I used a model-free estimator based on causal discovery.

In **Chapter** 2 I used a type of neural network called a restricted Boltzmann machine to estimate interactions in various settings. To establish a baseline, **Section** 2.3.1 shows estimates of interactions in simulations of Ising models where the ground truth was known. I first reproduced the results on pairwise Ising interactions from [59], and then moved on to Ising models with third-order interactions. To explore the interactions in a context where the ground truth was not known, **Section** 2.3.2 shows a side project where RBMs were trained on population-level trait data from the UK Biobank. With the developed intuition in hand, I binarised a developmental dataset of astrocytes, and calculated interactions among a set of astrogliogenesis genes in **Section** 2.3.3. There were various problems with the estimates, which are reflected upon in the Discussion in **Section** 2.4, that made it impossible to interpret the inferred interactions.

To address these issues, in **Chapter** 3 a model-free estimation approach is outlined. It starts by showing how model-bias arises in **Section** 3.1.1, and in **Section** 3.2.1 shows how the model-free approach is defined and how it evades this problem. **Section** 3.2.2 introduces the practical aspects involved in the model-free estimation procedure and **Section** 3.2.3 introduces different ways to validate the estimates.

Section 3.3 shows a number of different results: **Section** 3.3.1 repeats a theoretical result that previously appeared as part of [125] and which explicitly links the MFIs to information theory, and **Sections 3.3.2 and 3.3.3** compare higher-order interactions and information theoretical quantities in logic gates and other causal structures.

Chapter 4 explores the mechanistic interpretation of MFIs. In **Section** 4.3.1, I for the first time calculated MFIs on gene expression data, and explored how stable, reproducible, and robust the estimates were for different numbers of cells and genes. Various results regarding the validation against known protein function and annotation are listed in the other subsections from **Section** 4.3.

In contrast to the mechanistic interpretation, and more fruitfully, **Chapter** 5 explores the cell states implied by up to fifth order interactions. How such states are defined is introduced in **Section** 5.2.1, and results in mouse neurons and astrocytes are discussed in **Section** 5.3.

Finally, **Chapter** 6 concludes the thesis and discusses some limitations and possible improvements to the estimation method.

It should be noted that this thesis is the result of a very diverse and transdisciplinary research project. While I have tried to be *adisciplinary* and not separate the chapters into different disciplinary perspectives, readers from different backgrounds will inevitably be drawn to different chapters and sections. For example, as I did not end up using the RBM estimation to answer any of the central questions in this thesis, Chapter 2 will mostly be interesting to readers specifically interested in the learning and sampling dynamics of RBMs *per se*. While Chapter 3 offers an important introduction to MFIs, readers not interested in their mathematical context or information theory should feel free to skip Section 3.3.1.

Chapter 2

Inferring interactions with restricted Boltzmann machines

A[h], la recherche! Du temps perdu.

Marcel Proust [200] (punctuation mine)

䷃

2.1 Introduction

This chapter concerns the estimation of interactions with restricted Boltzmann machines, and serves as an introduction to the maximum entropy interactions in genetic data. The reason that RBMs are an obvious choice for maximum entropy estimation is that they are closely related to Ising models. They are, in a sense, trainable Ising models. This similarity is immediately obvious from their definition, and that they can encode the data as Ising models was already anticipated in *e.g.* [178], but a precise mapping from RBMs to Ising models was first introduced in [59]. The results from [59] serve as the starting point for this chapter. To train the RBMs with the contrastive divergence algorithm, I used the `pytorch` implementation from the authors of [59], but I implemented the persistent contrastive divergence and parallel tempering training schemes myself.

While RBMs are thus closely related to statistical models from physics, they also naturally arise in a purely statistical context, which is how I will introduce them in this thesis, to emphasise that their interpretation is not contingent on the physical meaning and justification of the Ising model.

2.1.1 Boltzmann machines and couplings

The Ising model as it was introduced in the introduction is an example of a more general class of models: Markov Random Fields. A Markov random field is a set of random variables V with an associated graph $G = (V, E)$ such that the following property holds[1]:

$$p(V) = \prod_{C \in \mathfrak{C}(G)} \phi_C(C) \tag{2.1}$$

where $\mathfrak{C}(G)$ is a set of cliques, *i.e.* complete subgraphs, in G. That is, the joint probability distribution over all variables V factorises into a product of so-called *clique potentials* ϕ_C of the cliques in G. It is immediately clear that the Ising model is a special case: the cliques are the nearest neighbours on a lattice, and the clique potentials are just the (unnormalised) marginal Gibbs weights:

$$P(X) \propto \prod_{(X_i, X_j) \in \mathfrak{C}(G)} \phi_{ij}(X_i, X_j) \tag{2.2}$$

where

$$\phi_{ij} = \exp\left(-JX_iX_j - h(X_i + X_j)\right) \tag{2.3}$$

However, the cliques do not in general have to be so small, or regular. For the fully connected graph, one choice of cliques is all pairs $(X_i, X_j) \in X \times X$, which leads to clique potentials

[1] This assumes that $p(V)$ has full support, in which case it follows from the Hammersley-Clifford theorem. Markov random fields can be defined in a more general setting, but that is beyond the scope of this thesis.

$$\phi_{ij} = \exp\left(-J_{ij}X_iX_j - h_iX_i - h_jX_j\right) \tag{2.4}$$

The Markov random field with these clique potentials is called a *spin glass* model, but may also be interpreted as an artificial neural network where neuron X_i updates its state according to the following activation function [108]:

$$p(X_i = 1) = (1 + \exp(-h_i + \sum_j J_{ij}X_j)) \tag{2.5}$$

This neural network, called a Boltzmann machine (BM), can be used to solve both the forward, and the inverse problem. To solve the forward problem of predicting the dynamics from a certain set of weights, the signals can be propagated through the network, and the equilibrium dynamics will approximate the joint distribution $p(X)$. To solve the inverse problem of inferring the interactions that produce a particular distribution, the biases h_i and weights J_{ij} are changed iteratively in a training procedure. Denoting the expected value across the data distribution and the model distribution with $\langle\ ,\ \rangle_{\text{data}}$ and $\langle\ ,\ \rangle_{\text{BM}}$ respectively, the gradient with respect to the weights J_{ij} of the log-likelihood on observed data Y is

$$\left\langle \frac{\partial \log p(X)}{\partial J_{ij}} \right\rangle_{\text{data}} = -\left\langle \frac{\partial E(X)}{\partial J_{ij}} \right\rangle_{\text{data}} - \left\langle \frac{\partial \log \mathcal{Z}}{\partial J_{ij}} \right\rangle_{\text{data}} \tag{2.6}$$

where $E(X) = \sum_{ij} J_{ij}X_iX_j + h_iX_i + h_jX_j$, and the normalisation $\mathcal{Z} = \sum_{X\in\mathcal{X}} \exp(-E(X))$.

$$= \langle X_iX_j\rangle_{\text{data}} - \left\langle \frac{1}{\mathcal{Z}} \frac{\partial \mathcal{Z}}{\partial J_{ij}} \right\rangle_{\text{data}} \tag{2.7}$$

$$= \langle X_iX_j\rangle_{\text{data}} - \left\langle \frac{1}{\mathcal{Z}} \sum_{X\in\mathcal{X}} X_iX_j \ \exp(-E(X)) \right\rangle_{\text{data}} \tag{2.8}$$

$$= \langle X_iX_j\rangle_{\text{data}} - \left\langle \langle X_iX_j\rangle_{\text{BM}} \right\rangle_{\text{data}} \tag{2.9}$$

$$= \langle X_iX_j\rangle_{\text{data}} - \langle X_iX_j\rangle_{\text{BM}} \tag{2.10}$$

and similarly for the biases. By updating the weights and biases according to this training step, the BM can approximate the data distribution in the dynamics of its nodes. In fact, given enough data, this training procedure is convex [108]. However, the BM can only model pairwise interactions, which are in general not sufficient [163, 22]. To allow higher-order couplings to appear, a new set of *hidden* nodes h is added to the BM, while the old variables are now referred to as the *visible* nodes v.

To get an estimate of $\langle v_ih_j\rangle_{\text{data}}$ or $\langle h_ih_j\rangle_{\text{data}}$, the visible nodes are set to a sample from the training data, and the hidden nodes are sampled until they equilibrate, for which there is no efficient algorithm [108]. Reaching equilibrium is made difficult by the connections between the hidden nodes, so one is naturally led to the situation where the graph is bipartite, and there are no connections among visible nodes or among hidden nodes. A Boltzmann machine with this bipartite structure is called a **restricted Boltzmann**

machine, or RBM. In contrast to fully connected BMs, RBMs can be efficiently trained, as outlined in Section 2.2.1. Since an RBM only has connections between visible and hidden nodes, the probability distribution over the states $s = (v, h)$ of a machine with n visible and m hidden nodes can be written as

$$p(v,h) = \frac{1}{\mathcal{Z}} e^{-E(v,h)} \tag{2.11}$$

with

$$E(v,h) = -\sum_{i}^{n}\sum_{j}^{m} (v_i w_{ij} h_j + b_i v_i + c_j h_j) \tag{2.12}$$

where w_{ij} is the weight of the connection between visible node i and hidden node j, and b_i and c_i are the biases of the visible and hidden nodes, respectively. The visible layer forms an interface between the data and the machine given by the marginal probability distribution over the visible nodes:

$$\begin{aligned} p(v) &= \sum_{h \in \mathbb{B}^m} p(v,h) && (2.13) \\ &= \frac{1}{\mathcal{Z}} \prod_{i=1}^{n} e^{b_i v_i} \prod_{j=1}^{m} \left(1 + e^{c_j + v_i w_{ij}}\right) && (2.14) \end{aligned}$$

Writing the exponent as a power series reveals that this distribution encodes an infinite series of interactions between different powers of the visible nodes. However, on binary variables, $v_i^n = v_i \ \forall n$, so the infinite series of polynomial interactions reduces to all possible multilinear interactions between the variables, the strength of which is fully encoded in the weights and biases of the RBM. This gives it the structure of a generalised Ising model, and in [59] and [24], the authors explicitly show how to extract the interactions from the network weights. For example, the pairwise interaction between visible nodes v_{j_1} and v_{j_2} is

$$J_{j_1 j_2} = \frac{1}{8} \sum_i \frac{(1 + e^{c_i + w_{ij_1} + w_{ij_2}})(1 + e^{c_i})}{(1 + e^{c_i + w_{ij_1}})(1 + e^{c_i + w_{ij_2}})} \tag{2.15}$$

This means that when the hidden layer is sufficiently large to encode all dependencies in the data, the RBM is a *universal approximator* for distributions over binary variables [84, 169]—it can approximate any data distribution arbitrarily well.

2.1.2 Aim and outline of this chapter

In this chapter, I used Equation (2.15), and the corresponding expression for the 3-point interaction, to extract the interactions from trained RBMs and thus fit maximum entropy interactions to the data.

The results from this chapter provided insight into the estimation of maximum entropy interactions using RBMs, but ultimately mainly served to justify the model-free approach in Chapter 3.

This chapter starts by outlining a training procedure for RBMs called contrastive divergence in **Section 2.2.1**. A refinement of this, called parallel tempering, is introduced in **Section 2.2.2**. To be able to validate the trained RBMs against a known ground truth, **Section 2.2.3** introduces a method for simulating Ising models with third-order interactions. **Section 2.2.5** introduces the gene expression data from developing murine astrocytes.

I trained RBMs on three very different kinds of data: simulated Ising models (results in **Section 2.3.1**), epidemiological data from the UK Biobank (**Section 2.3.2**), and gene expression data from murine astrocytes (**Section 2.3.3**). The RBMs were able to capture the training distributions well, but there were some technical issues which make the interactions hard to interpret, as discussed in **Section 2.4**.

2.2 Methods

2.2.1 Training RBMs

When training RBMs, several hyperparameters need to be set. A list of these is reported in Table 2.1.

Parameter	Role
n_{vis}	No. of visible nodes, dimensionality of input data
n_{hid}	No. of hidden nodes
Epochs	No. of times training data should be seen
Batch size	No. of training examples to average the gradient over
α	Learning rate
k_{CD}	Steps in Contrastive Divergence chain
N_{PT}	No. of parallel tempering chains (optional)

Table 2.1: Hyperparameters of RBM training.

Maximising the log-likelihood in Equation (2.6) minimises the KL-divergence between the marginal distribution over the visible nodes and the data distribution. I used stochastic gradient descent (SGD) to minimise the negative log-likelihood by dividing the full data set in multiple batches of a fixed batch size, and calculating the gradient on each batch separately, each time changing the parameters by a step size α, a parameter known as the learning rate. I trained the machines by iterating over the full data set N times, where N is referred to as the number of epochs. However, calculating the gradient of the log-likelihood along the weights and biases requires taking expectation values with respect to the model distribution, which involves an intractable sum of all possible RBM states. Instead, I sampled states from the RBMs distribution and approximated the true gradient with expectation values with respect to the sampled states. Note that both layers in the

RBM induce a conditional distribution over the other one: $p(v|h, \theta = \{w, b, c\})$ and $p(h|v, \theta)$, and that within a layer all nodes are conditionally independent. All nodes in a layer can thus be sampled simultaneously in a process known as Gibbs sampling. In [107], Hinton showed that a Gibbs sampling procedure known as the Contrastive Divergence (CD) algorithm can yield an efficient and surprisingly accurate approximation to the gradient of the likelihood. The CD algorithm starts by setting the visible layer to a state from the current training batch, and then propagates it back-and-forth through the two layers k_{CD} times by Gibbs sampling from the conditional distributions, to end up with a (biased) sample from the model distribution [50]. To monitor training progress, there are a few metrics that can be calculated as training progresses. The most obvious is the KL-divergence between the training distribution and the marginal distribution over the visible nodes of the RBM. However, as shown in Appendix 2.C, this KL-divergence offers no more information than the log-likelihood of the data given the RBM's model, so I focused on the log-likelihood throughout this chapter. The log-likelihood of a particular batch S can be written as

$$\mathcal{L}(\mathcal{S}) = \sum_{s \in S} \log p(s) = \sum_{s \in S} \log \sum_{h \in \mathbb{B}^n} e^{-E(s,h)} - \log \mathcal{Z} \tag{2.16}$$

While it is too computationally expensive to calculate the log-likelihood for each training step, it can be estimated every few epochs using annealed importance sampling (see [59] for more details). Furthermore, as the log-likelihood gets maximised, the partition function approaches a constant, which means that the free energy, defined as

$$F(v) = \log \sum_{h \in \mathbb{B}^n} e^{-E(v,h)} \tag{2.17}$$

should also approach a constant, which can be used to monitor training progress. Additionally, since RBMs are generative models, samples from their visible layer can be compared with the training data. To do so, define the *magnetisation* (the first moment) of a state v as

$$m^{(v)} = \frac{1}{n} \sum_{i=1}^{n} v_i \tag{2.18}$$

The nth moment of a set S of observations is then defined as

$$\mu_n = \left\langle \Big(m - \langle m \rangle\Big)^n \right\rangle \tag{2.19}$$

$$= \frac{1}{N} \sum_{j=1}^{N} (m^{(j)} - \frac{1}{N} \sum_{k=1}^{N} (m^{(k)}))^n \tag{2.20}$$

where $N = |S|$. To monitor and evaluate the RBMs during training, I compared moments of the training data with moments of generated samples from the RBMs. A probability distribution uniquely determines its moments, and if the distribution is reasonably well-behaved, *i.e.* has a moment-generating function, the moments uniquely determine the distribution, so a well-trained RBM should generate samples with the same moments as the training data.

2.2.2 Parallel tempering

Contrastive divergence training suffers from the fact that the Markov chains converge to the true distribution more slowly as the network weights grow, making training unstable [82]. To stabilise training, I explored a technique called parallel tempering. One problem with contrastive divergence is that the gradients are only calculated closely around training examples. To get around this, one could increase k_{CD}, but doing so quickly becomes computationally expensive. Another solution, referred to as *persistent* contrastive divergence (PCD), is to first initialise the Markov chain with a random state, but afterwards use the sample that resulted from k_{CD} Gibbs samplings to initialise the next chain. Like this, the gradients can be estimated in a less restricted part of sample space. However, since the network weights change at each PCD iteration, the Markov chain is never actually initialised with a sample from the model distribution. To mitigate this effect, PCD requires a smaller learning rate [82]. However, PCD still suffers from emphasising areas of model space where the chain is currently running. Parallel tempering is a technique designed to explore a larger area of the model space of the RBM. It does so by running multiple Markov chains at the same time, all running at different temperatures. The distribution over an RBM

$$p(v, h|\theta) = \frac{1}{\mathcal{Z}(\theta)} e^{E(v,h|\theta)} \tag{2.21}$$

is a Boltzmann weight at unit temperature. Consider a family of K distributions, indexed by inverse temperatures $\beta_k = \frac{1}{T_k}$, $k \in [1, ..., K]$:

$$p_k(v, h|\theta) = \frac{1}{\mathcal{Z}_k(\theta)} e^{\beta_k E(v,h|\theta)} \tag{2.22}$$

Note that $\beta = 0$ corresponds to infinite temperature: a uniform distribution. Any set $\{\beta_k\}$ with βs between 0 and 1 interpolates between the original RBM distribution $p_{k=1}(v, h|\theta)$ and the uniform distribution $p_{k=0}(v, h|\theta)$. Parallel tempering then starts K parallel Markov chains, all at different temperatures. Only the $\beta = 1$ chain still corresponds to the original distribution, while the chains with lower β (higher temperature) explore the RBM's state space more uniformly. The question is how to use this information to inform the chain running at $\beta = 1$. The parallel tempering algorithm lets the K chains run independently for several steps, and then swaps the state of the two neighbouring chains running at temperature β_k and β_{k+1} with a probability given by a metropolis factor:

$$P(\text{swap}) = \min\left\{1,\ \exp\Big((\beta_k - \beta_{k+1})\left(E(v_k, h_k|\theta) - E(v_{k+1}, h_{k+1}|\theta)\right)\Big)\right\} \tag{2.23}$$

where v_k, h_k are the states from a chain running at temperature β_k. Then, like with PCD, after each iteration the final state for *each* of the K chains is used as the input to the next iteration. Parallel tempering is thus the same as running K PCD chains. To calculate expectation values, only states from the $\beta = 1$ chain are used, the difference with PCD being that this chain is now enriched with states that were sampled from more uniform distributions but still had a high probability under the $\beta = 1$ distribution. The only thing left to specify is how to do the state-swapping as the order in which the chains are compared might change results. Throughout this chapter, I followed the literature [68] and used the Deterministic Even Odd (DEO) algorithm, first proposing swaps between all chains at β_k and β_{k+1} for even k, and then starting again at odd k.

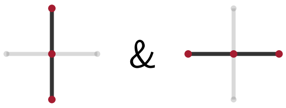
2.2.3 Simulating Ising models

I used the `c++` program `Magneto` [296] to simulate the equilibrium dynamics of a homogeneous Ising model with nearest-neighbour couplings, doubly periodic boundary conditions (*i.e.* on a torus), in the $\{-1, 1\}$ basis. I set the Boltzmann constant $k_B = 1$ throughout this thesis so that the only two physical parameters were the temperature and the coupling strength. Sampling was done using the metropolis algorithm which requires an expression for the energy difference that results from a spin-flip. The effect of a spin flip for the nearest neighbour coupling structure was already implemented in `Magneto`, but not for systems with a three-point coupling but without pairwise nearest neighbour interaction or external fields, so I implemented this myself. The Boolean variables could be represented as $\{0, 1\}$ or $\{-1, 1\}$, and these two bases should be carefully distinguished. Consider a system that has only a 3-point interaction in the $\{0, 1\}$ basis. Its Hamiltonian, or energy function, can be written as follows:

$$\mathcal{H} = \sum_{i,j,k} J_{ijk} \; v_i v_j v_k \Big|_{\{0,1\}} \tag{2.24}$$

$$= \sum_{\langle ijk \rangle} J^{(3)} v_i v_j v_k \Big|_{\{0,1\}} \tag{2.25}$$

Where the sum in the second line only includes triplets that lie in a straight connected line, and the restriction indicates the basis. That is, only couplings of the following type appear:



Relating this to `Magneto`'s $\{-1, 1\}$ basis by sending $v \to \frac{v+1}{2}$, the Hamiltonian becomes:

$$\mathcal{H} = \sum_{i,j,k} J_{ijk} \left(\frac{v_i + 1}{2}\right)\left(\frac{v_j + 1}{2}\right)\left(\frac{v_k + 1}{2}\right)\Big|_{\{-1,1\}} \tag{2.26}$$

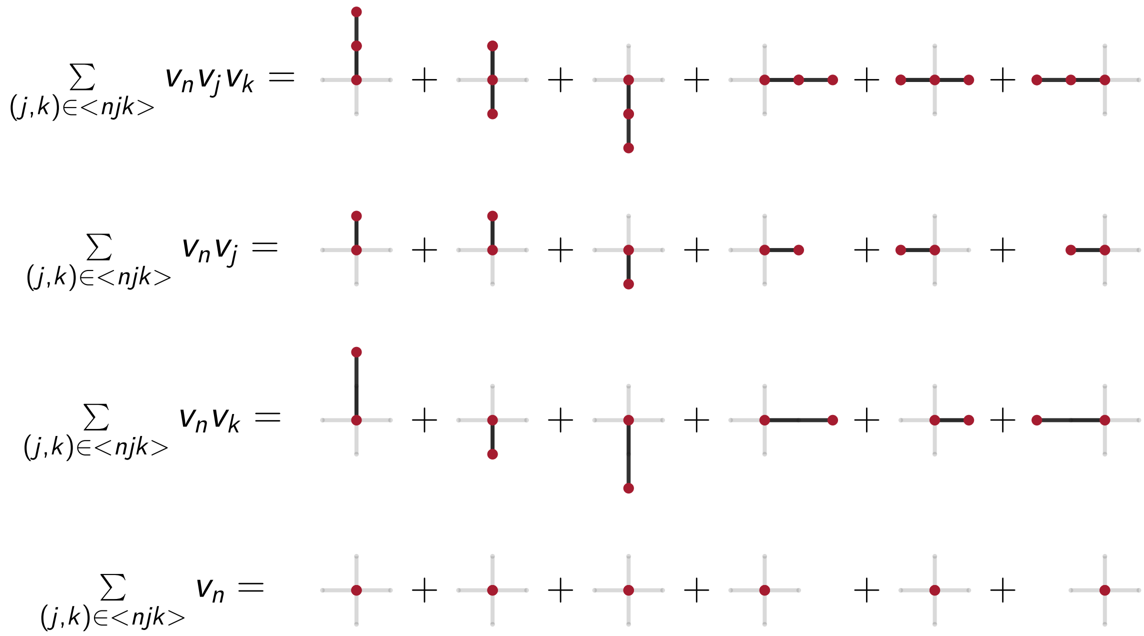
$$= \sum_{i,j,k} \frac{J_{ijk}}{8} \left(v_i v_j v_k + v_i v_j + v_i v_k + v_j v_k + v_i + v_j + v_k\right)\Big|_{\{-1,1\}} + \text{constant} \tag{2.27}$$

Now, consider the change in energy resulting from a spin flip at site n. The full calculation is presented in Appendix 2.A and results in:

$$\Delta E_n = \mathcal{H}_{v_n \to -v_n} - \mathcal{H} \tag{2.28}$$

$$= -2 \sum_{(j,k) \in <njk>} \frac{3 \cdot J^{(3)}}{8} \Big(v_n v_j v_k + v_n v_j + v_n v_k + v_n \Big) \tag{2.29}$$

where the sum includes all pairs of sites (j,k) that form a straight connected triplet with site n. Order by order, the terms that appear in ΔE_n can be represented diagrammatically as:


$\sum_{(j,k) \in <njk>} v_n v_j v_k =$ + + + + +
$\sum_{(j,k) \in <njk>} v_n v_j =$ + + + + +
$\sum_{(j,k) \in <njk>} v_n v_k =$ + + + + +
$\sum_{(j,k) \in <njk>} v_n =$ + + + + +

A spin-flip in a *Magneto* simulation should thus affect:

- Six 3-point couplings with weight $3 \cdot J^{(3)}/8$
- Four nearest neighbour couplings with weight $2 \cdot 3 \cdot J^{(3)}/8$
- Four next-to nearest neighbour couplings with weight $3 \cdot J^{(3)}/8$
- One linear term with weight $6 \cdot 3 \cdot J^{(3)}/8$

which I implemented as an alternative energy function in `magneto`, and used throughout the rest of this chapter.

Note that there is a difference between the triplet coupling $J^{(3)}$ which appears when the sum contains only connected triplets, and the coupling tensor J_{ijk} that appears when the sum contains all triplets. For a given spin flip at site n, J_{njk} has two-fold degeneracy in j and k, while $\sum_{(j,k) \in \langle njk \rangle} J^{(3)} v_n v_j v_k$ covers six different triplets. This implies that the two are related through

$$\frac{J^{(3)}}{2} = \frac{J_{ijk}}{3!} \tag{2.30}$$

and the RBMs should thus encode a triplet interaction with coupling strength $\frac{J^{(3)}}{3T}$, where T is the temperature of the simulated system.

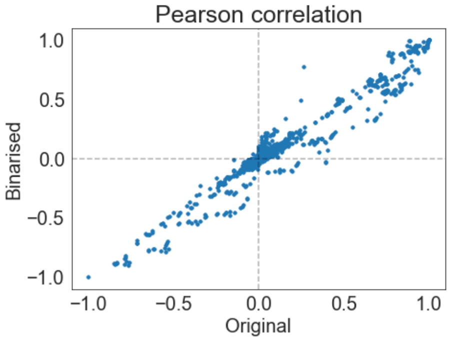


Figure 2.1: The Pearson correlation between each of the UK Biobank traits is mostly robust to binarisation.

2.2.4 UK Biobank traits

The UK Biobank is a data set with phenotypic and genotypic data from over 500,000 individuals [256]. Many of these phenotypes correspond to disease phenotypes that can be represented as a binary variable indicating the presence or absence of the corresponding diagnosis. Other phenotypes, like blood counts or body weight, have to be binarised before they can be used as inputs to an RBM. In this section, and Section 2.3.2, the binarisation of the data and training of the machines has been performed by Ava Khamseh, who had access to individual-level data in the UK Biobank. We used data from 400,000 individuals, and selected 62 traits that had been of interest to my colleague Neil Clark (in unpublished work and private communication). In addition, we added two traits corresponding to male or female sex, and three traits corresponding to age bins of $[40, 49], [50, 59]$, and $[60, 69]$ years. Of these 67 traits, 26 were already binary, and the other 41 traits were binarised around their median value in the cohort. A full list of traits is printed in Appendix 2.E. The Pearson correlation between each of the traits was mostly robust to binarisation (see Figure 2.1).

2.2.5 Astrocyte gene expression

2.2.5.a 10X data generation

The Million Cell Data Set (MCD) [4] from *10X Genomics* comprises 1,306,127 transcriptomes of cells from the cortex, hippocampus and ventricular zone of two E18.5 mouse brains. The brain tissues were dissociated using *10X Genomics*'s own protocol [1], and 133 barcoded cDNA libraries were created using the *10X Genomics* Single Cell 3‘ v2 chemistry kit, aiming to include around 10,000 cells per library. These 133 libraries were then processed using 17 *Chromium* chips, each of which can process 8 libraries in parallel. The libraries were sequenced on an Illumina HiSeq 4000, using paired-end sequencing, at a moderate read depth of around 18,500 reads per cell. During library preparation, each RNA molecule got tagged with a unique molecular identifier (UMI), so there was no ambiguity between reads, fragments, and transcripts, and all reads could be directly converted to transcripts per million (TPM) without correcting for gene length or PCR bias. This leads to TPM values for 27,998 genes across 1,306,127 transcriptomes, with a median of 1,870 expressed genes and 5,000 transcripts per cell, keeping only uniquely mapped reads. Alignment batch correction due to mouse or library, and

Louvain-clustering were performed by *10X Genomics* with their in-house `CellRanger` software.

2.2.5.b Removing doublets

The *Chromium* system is droplet-based and aims to suspend each cell in its own oil droplet. However, as more cells need to be barcoded, the chip gets loaded with more cells which increases the probability of two cells ending up in the same droplet, forming a *doublet*. Since each library in the MCD contained around 10k transcriptomes, the expected doublet rate was relatively high at 7.6% [2]. Doublets result in transcriptomes that do not correspond to any cell, so can distort the result of any downstream analysis and should be removed. I used the Python package `Scrublet` [302] to identify and remove potential doublets. I first log-transformed the normalised UMI counts, and only kept overdispersed genes (using the `scanpy` function `scanpy.pp.highly_variable_genes`). To detect doublets, `Scrublet` creates artificial doublet transcriptomes by combining transcriptomes from different clusters. These artificial doublets form new clusters, and real transcriptomes can be given a doublet score by how close they are to a cluster of artificial doublets. Some simulated doublets separate from the observed transcriptomes and form *neotypic* doublet transcriptomes, while others fall within the existing clusters, forming *embedded* doublet transcriptomes. I ran the `Scrublet` algorithm on the full data set, embedding each transcriptome in the first 30 principal components, based on the top 15% most variable genes. The number of simulated doublets was 0.5 times the total number of transcriptomes. The number of nearest neighbours K in the KNN-graph was automatically set to $K = \text{round}(0.5 \times \sqrt{n_{\text{cells}}})$. The distribution of simulated doublet scores is shown in Figure 2.2. The two modes of the simulated doublets, corresponding to the embedded and neotypic doublets, were clearly visible, and the automatically set threshold at 0.34 separated the two peaks well. At this threshold, the total number of detected doublets was $53,048$, or 4.1% of all cells. The estimated proportion of detectable doublets was 48.7%, leading to a predicted total doublet rate of 8.3%, slightly higher than the 7.6% predicted in the documentation of the *10X* protocol. That the histogram of simulated doublet scores was so strongly bimodal should be considered evidence that the threshold at 0.34 was robust and appropriate. I therefore removed all cells annotated as doublets from all further analysis.

2.2.5.c Quality control

A UMAP and PCA embedding of all cells is shown in Figure 2.4. I performed a standard QC on the data: normalising expression data per cell, removing droplets with fewer than 600 detected genes and removing droplets with more than 12% of reads mapping to mitochondrial genes (see Figure 2.3) . Batch effects due to mouse or library were already removed by *10X Genomics*. For downstream analysis, highly variable genes were identified with the `Python` package `Scanpy` (using again the `scanpy.pp.highly_variable_genes` function).

2.2.5.d Cell type annotation

The `R` package `SingleR` was used to annotate the cells according to cell type. It contains a reference data set of 358 annotated mouse samples to compare with. However,

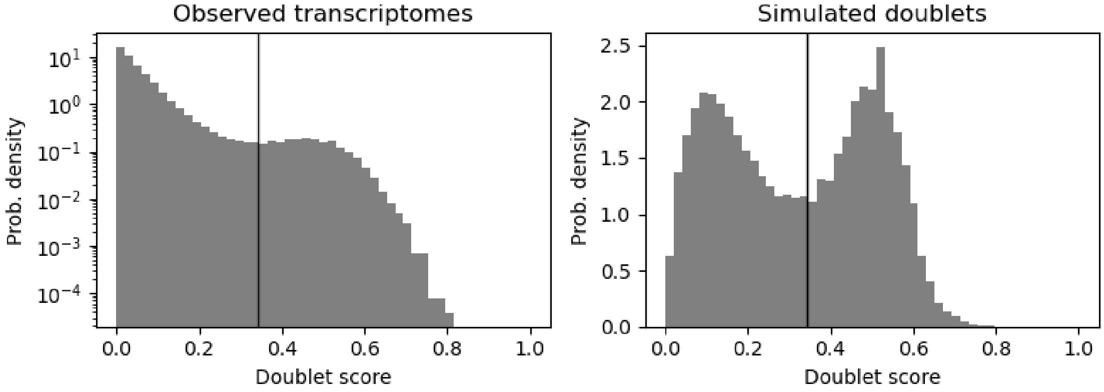


Figure 2.2: The distribution of simulated doublets (right panel) is strongly bimodal, indicating successful placement of the doublet-score threshold. This threshold led to a doublet rate of 8.3% across all 1.3M observed transcriptomes (left panel).

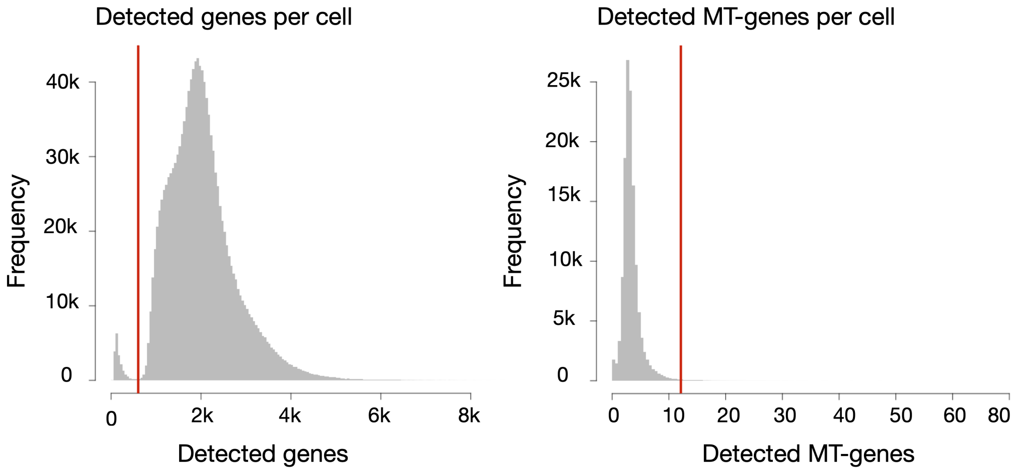


Figure 2.3: QC metrics across all 1.3M cells, with the cut-off line in red. Droplets with fewer than 600 detected genes or more than 12% of mitochondrial reads were removed.

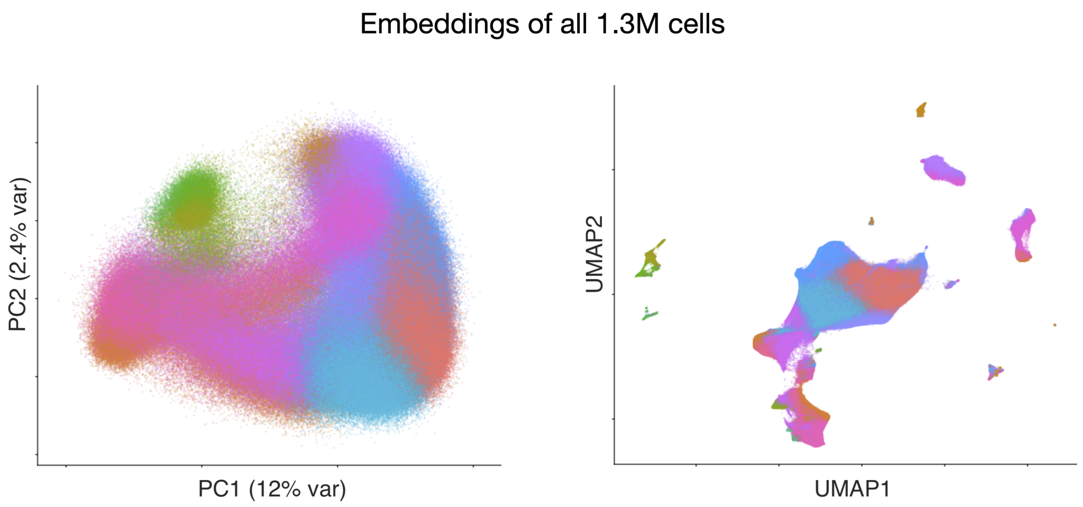


Figure 2.4: Embeddings of all 1.3M cells. **Left**: PCA, **right**: UMAP. Coloured by the Louvain cluster identity that *10X Genomics* provided. The clusters were indeed connected in the embeddings, and there is clear structure present in the data.

since these are normalised bulk expression data instead of single-cell samples, it is recommended to add single-cell markers from the literature [13]. I focused on astrocytes, since *10X Genomics* claimed that they formed the largest non-neural population in their data set. Based on results from [267], [43], and [313], I added the following genes as astrocyte markers: *F3, Rorb, Acsbg1, Ntsr2, Plcd4, Gja1, Gjb6, Cbs, Chrdl1, Prodh, Mlc1, Acsl6, Slc4a4, Gabrg1, Cxcl14, Slco1c1, Vcam1, Ednrb, Scrg1, Bcan, Aldoc, Gfap, Aldh1l1, Aqp4, Serpinf1, Mfge8*. In total, 43,966 cells got annotated as astrocytes (i.e. 3.5% of all cells), almost all of which were part of cluster 8 of the k=20-means clustering (see Figure 2.5). This is significantly fewer astrocytes than *10X Genomics'* own analysis, where they identified 14.4% of all cells as Astrocytes [167], which probably means that cluster 8 corresponds to one of multiple astrocyte clusters. However, it is clear from Figure 2.5 that the 43,966 cells formed a relatively homogeneous set. Since there could be different transcriptionally distinct subtypes or states of astrocytes present in the data, I used cluster 8 as the training data, regardless of `SingleR` annotation. In total, this cluster contained 77,524 cells.

2.2.5.e Gene selection

Inspired by the idea of *core genes* in a model of omnigenic trait inheritance [38], I chose to focus on transcription factors as an interesting set of genes potentially mediating the interactions. Of particular interest were genes relevant to the late developmental astrocytes that are the focus of this study. The authors of [270] considered the gene expression profile of E18.5 mice as neural stem cells differentiate into astrocytes, a process known as astrogliogenesis. Having identified in particular the transcription factors (TFs) Nfia and Atf3 as playing a role in astrogliogenesis, they suppressed the expression of these two TFs (by siRNA-mediated depletion of the corresponding transcripts), and looked for differentially expressed genes. To enrich the set of differentially expressed genes for physical interactions, they filtered for genes that showed the enhancer activity

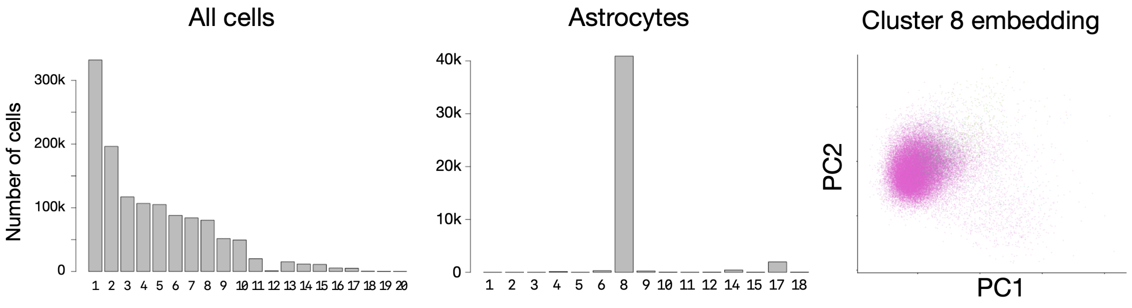


Figure 2.5: Histogram of the cluster identities of all cells that passed QC (left panel), of the astrocytes (middle panel) and a PCA embedding of the astrocytes (right panel). Almost all cells that are annotated as astrocytes by `SingleR` derive from cluster 8 of the k=20-means clustering provided by *10X Genomics*. It can also be seen that almost all cells in cluster 12 were removed as they are suspected to be doublets by `Scrublet`, and that the astrocytes form a relatively homogeneous set of cells.

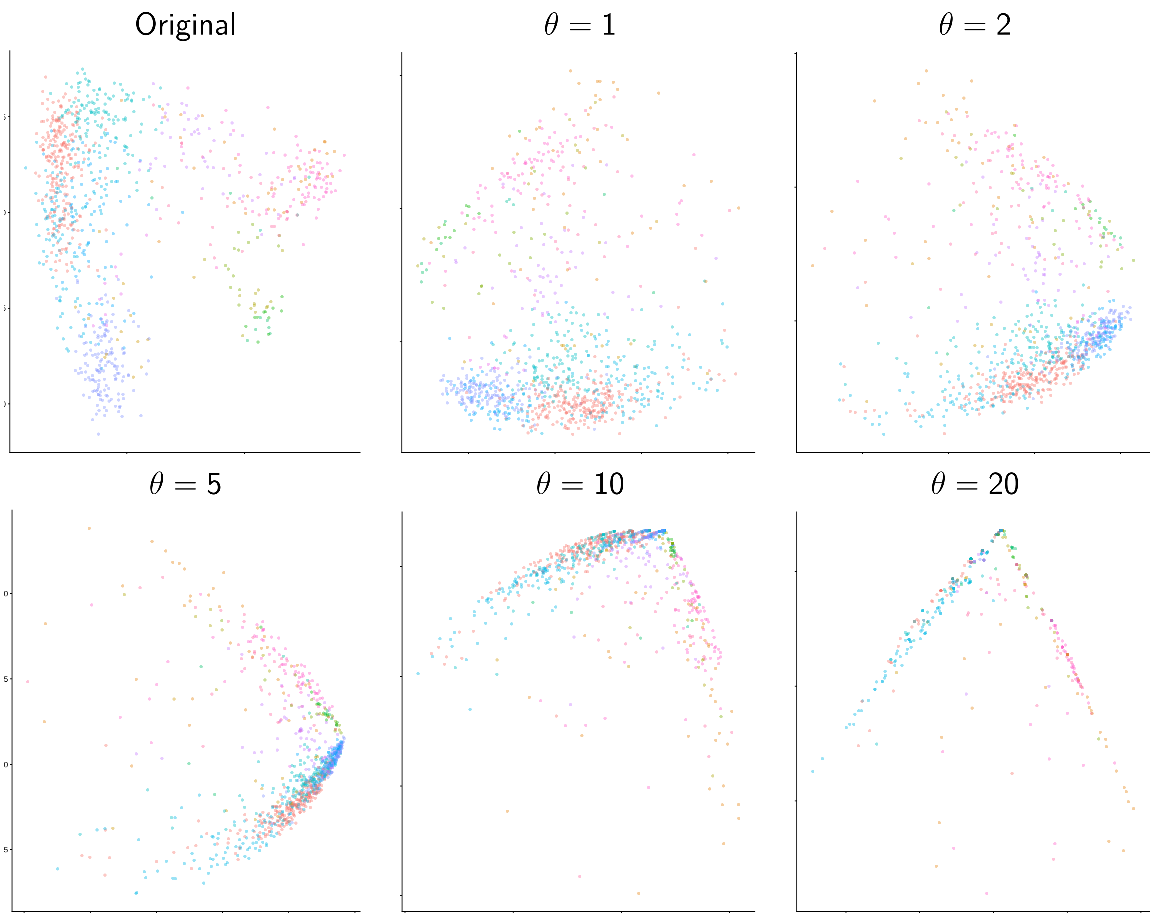


Figure 2.6: PCA embeddings of 20,000 randomly selected cells from the same data set shown in Figure 2.4, using the same clustering. The parameter θ represents the read count around which the expression values are binarised. Thresholding expression at increasing values of θ distorts the data, so that setting $\theta = 1$ preserves the most structure, and was the threshold used throughout this thesis.

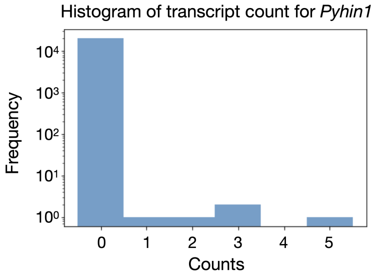


Figure 2.7: Across 4,000 randomly selected genes, only *Pyhin1* showed bimodal gene expression. However, the observed count frequencies are respectively $(19995, 1, 1, 2, 0, 1)$, which made the bimodality indistinguishable from noise.

marker H3K27ac, and were close to transcription factor binding motifs. The intersection of differentially expressed genes for Nfia and Atf3 that satisfied this demand comprised 47 genes. Additionally, to only keep genes significantly expressed in the cell population, only genes that are expressed in more than 1,000 cells, or around 2% of cells, were included in the final training data. This resulted in 37 genes of interest: *Creb1, Sned1, Nfasc, Atf3, Ggta1, Crb2, Atf2, Jun, Nfia, Fzd9, Hspb1, Cav2, Crtc3, Ampd3, Eef2k, Timp3, Ddit3, Crtc1, Ctsb, Slc39a14, Smad3, Slc38a3, Xbp1, Gas7, Cacng5, Phactr1, Pgf, Tgfb3, Ptp4a3, Atf4, Atf1, Clip4, Crem, Aqp4, Smad4, Fth1, Ppp1r3c*.

2.2.5.f Binarisation

To train RBMs on gene expression data and extract the interactions, the expression levels need to be binarised. The data is composed of read counts for every gene, so entries take values in $\mathbb{N}_0$. Binarisation therefore amounts to specifying a map $b : \mathbb{N}_0 \to \mathbb{B}$, where $\mathbb{B}$ is the Boolean domain $\{0, 1\}$. There are many ways to implement this map, and different binarisations could lead to different results. If a gene's expression is bimodal, then a natural choice of threshold would be one that separates the peaks of the distribution. Since most genes' distribution of expression has a mode at 0 reads, I looked for bimodality where one peak was at zero, and the second one was larger than 1. For 4,000 randomly selected genes I calculated the expression distribution across 20,000 randomly selected cells. I found just one gene—*Pyhin1*—with such a bimodal expression distribution. However, looking at the distribution of read counts in figure 2.7, the actual frequencies are so low that the bimodal signal seems indistinguishable from noise. I concluded that there is no evidence for bimodal gene expression.

Another obvious choice is thus thresholding at 1, which sends zero counts to the number 0, and every nonzero count to 1. Figure 2.6 shows that data thresholded at higher values than 1 showed increased distortion compared to the unbinarised data. One could consider more sophisticated maps, based on *e.g.* fitting Bernoulli models to scRNA-seq data as in [149], or scaling by library size, but the authors of [149] and [225] showed that thresholding at 1 works remarkably well, so throughout this thesis I implemented b as this threshold. That this simple approach worked well is perhaps not so surprising as scRNA-seq datasets contain many technical zeros. The technical noise of these experiments with only moderate read depth is so high, that a nonzero count does not reflect the transcript's concentration but rather the fact that the gene was expressed at all [204].

2.3 Results

This section describes three different contexts in which I trained RBMs and analysed the interactions. In **Section 2.3.1**, I simulated Ising models to compare the RBM interactions with a known ground truth. In **Section 2.3.2**, I analysed machines that were trained on various traits from the UK Biobank. The training of the machines, and the processing of the raw patient-level data was handled by Ava Khamseh, and all further analysis was done by myself. This data set bears little resemblance to gene expression data, but serves as an exploratory context in which to train RBMs on real data of clinical interest. Finally, **Section 2.3.3** presents the results on genetic interactions in embryonic astrocytes.

2.3.1 The RBMs reproduced the interactions in generalised Ising models

I first used RBMs to estimate interactions where the ground truth is known: a simulated (generalised) Ising model. I started with the canonical case of homogeneous, pairwise nearest-neighbour coupling on a square lattice with torus topology, *i.e.* double periodic boundary conditions, and then moved on to a model with only triplet interactions on the same toroidal lattice.

2.3.1.a Pairwise nearest-neighbour interactions

I started by reproducing the results from [59] on an 8×8 lattice with only a homogeneous nearest neighbour coupling. Using `Magneto`, I simulated 100k Ising states at $\mathrm{T} = 1.8$, after discarding the first 1k steps to let the MCMC chain thermalise. The training parameters are summarised in Table 2.2.

Epochs	Batch size	LR	k_{CD}	n_{vis}	n_{hid}
8k	200	0.1	1	64	64

Table 2.2: Training parameters for toroidal 8×8 Ising lattice at $T = 1.8$.

In Figure 2.8, the 1-, 2-, and 3-point interactions extracted from trained machines are shown. The 1-point interaction, or linear term, was accurately reproduced, and the 2-point interactions clearly showed the nearest-neighbour structure. The distribution of 3-point interactions sharply peaked around zero.

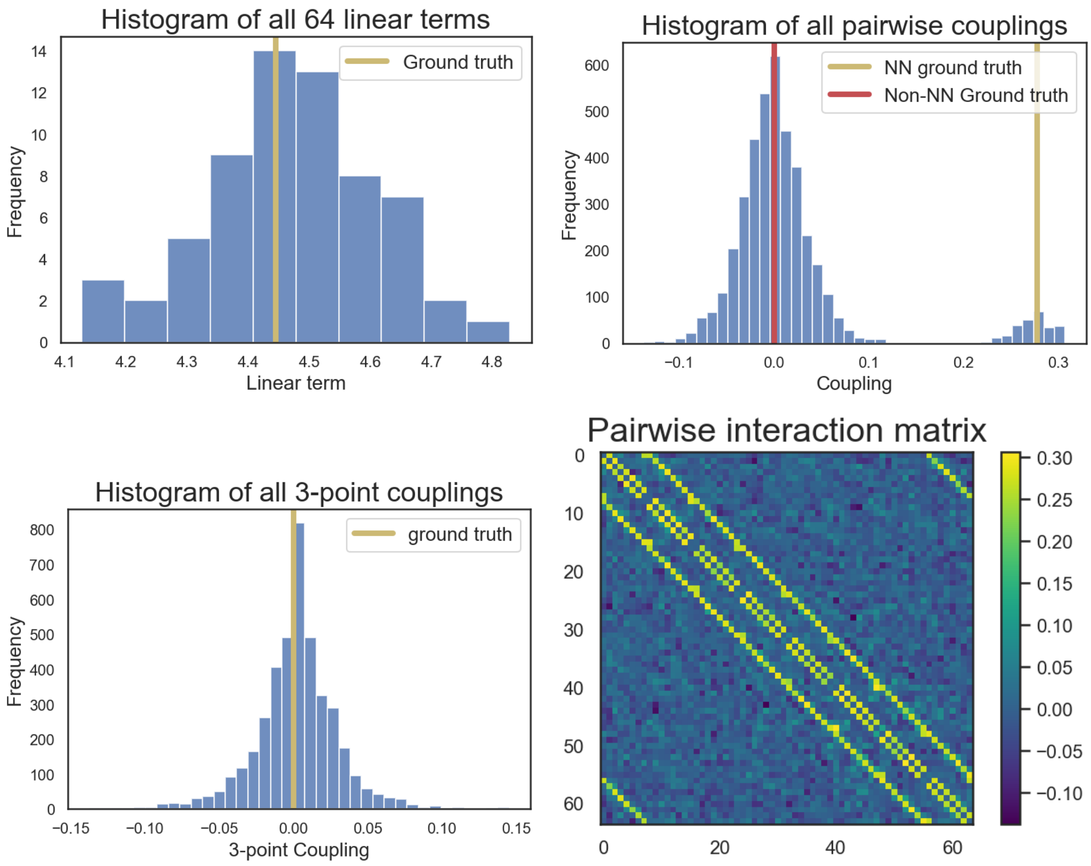


Figure 2.8: Couplings in the RBM's $\{0, 1\}$ basis at order 1, 2 and 3, extracted from a machine trained on 100k states from an Ising model with only nearest-neighbour (NN) interaction in the $\{-1, 1\}$ basis, at $T = 1.8$. At this temperature, the variables couple with a $\{0, 1\}$ nearest-neighbour coupling of $\frac{1}{3.6}$, a linear term of $\frac{8}{3.6}$, and no 3-point coupling, all of which were accurately reproduced by the trained RBMs.

2.3.1.b Variability across machines

While the overall performance of a trained RBM can be quantified in terms of the log-likelihood of the training data, the uncertainty in the estimates of the interactions is more difficult to quantify as there is no clear null hypothesis and there are different sources of variability. The training procedure is stochastic both in the initialisation of the weights, and in the batches from an instance of stochastic gradient descent. To quantify the variability as a result of this stochasticity, I trained 20 machines on the T=1.8, 8×8 Ising model data. All machines were trained with a learning rate of 0.1, a batch size of 200, for 8k epochs.

Figure 2.9 shows that the log-likelihood increased to a final value of -12.69 ± 0.98, leading to a KL divergence of 6.80 ± 0.98, in line with the $T = 1.8$ machine from [59] (reproduced in Figure 2.43 of this thesis).

The couplings that are encoded in the RBM are entirely determined by the weights of the network. The converse is not true: different weight matrices can lead to the same distribution (consider for example interchanging two hidden nodes and all their

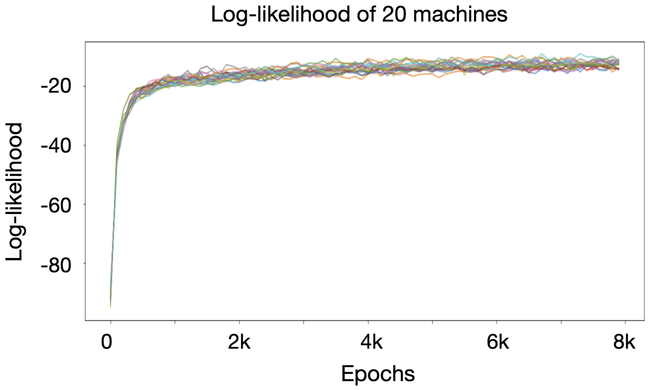


Figure 2.9: The log-likelihood of all 20 machines increased consistently and similarly during training.

n-pt	ground truth	mean	std
1-pt	4.44	4.48	0.16
2-pt (NN)	0.28	0.28	0.02
2-pt (non-NN)	0.00	0.00	0.03
3-pt	0.00	0.00	0.03
W	-	0.01	1.15
W^TW	-	4.50	8.89
v_{bias}	-	1.46	1.76

Table 2.3: The mean and standard deviation of the interactions and network parameters across the 20 RBMs. While the weights varied strongly across machines, the interactions were reproducible and accurate. The 2pt-interactions between nearest neighbours (NN) and non-nearest neighbours are presented separately.

connected weights). To study the variability across machines further, I compared the following quantities:

- The weight-matrix W of an RBM
- The bias of the visible nodes
- The matrix W^TW
- The encoded 2-point interactions

The three matrices are shown as heatmaps in Figure 2.10, and the variability in the 2-point couplings specifically is shown in Figure 2.11 as a violin plot.

In spite of the variability in the weights across the 20 machines, the RBMs were able to resolve the nearest neighbour structure and predicted the coupling strength with good numerical accuracy. Note that the variance in couplings across machines was smaller for nearest neighbours than it was for non-nearest-neighbours. This is summarised in Table 2.3, which shows the mean and standard deviation across the 20 machines for different orders of interactions and network parameters.

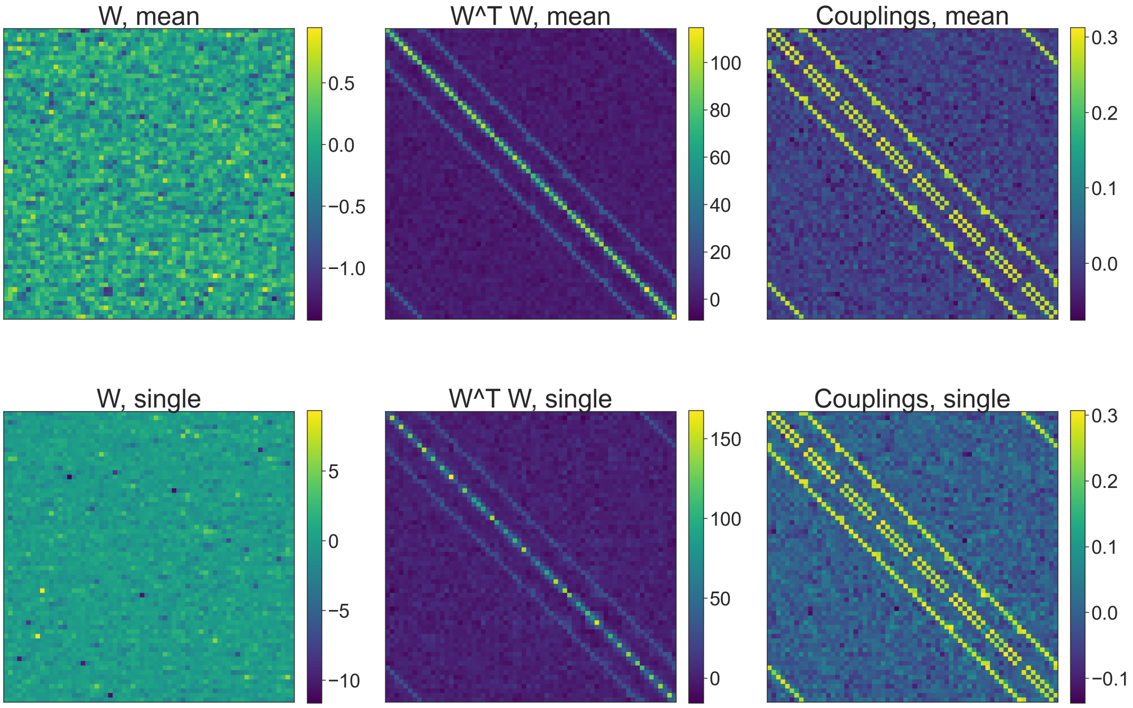


Figure 2.10: The weights and couplings encoded in the 20 trained machines. **Top:** Mean values across all 20 machines. **Bottom:** Values from a single machine. As expected, there was no consistent structure in the network weights W, but as soon as the hidden layer was 'summed out' in $W^T W$, the nearest neighbour structure appeared.

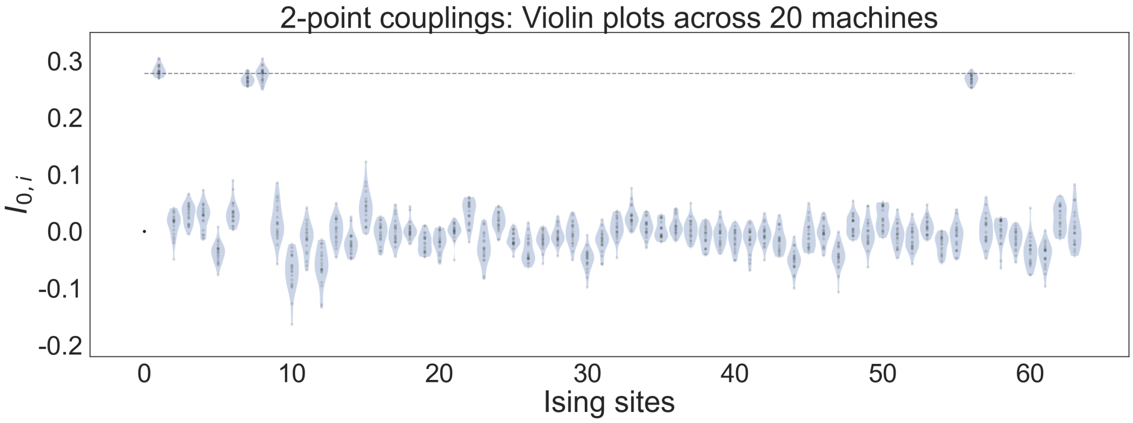


Figure 2.11: All interaction estimates for non-interacting pairs cluster around a value of zero, while the interacting pairs match the ground truth interaction, here indicated with a black dashed line. Note that the variance across machines is smallest for the interacting pairs.

2.3.1.c Triplet nearest-neighbour interactions

To expand on the results from [59], I simulated an Ising model with no linear or pairwise interactions, but a homogeneous 3-point interaction among triplets connected in a straight line on the toroidal lattice.

I generated 100k states from its Boltzmann distribution with $J^{(3)} = -1$ at temperatures from 1.5 to 2.5, and trained one machine on each of these data sets using the hyperparameters reported in Table 2.4. A range of values for each of the hyperparameters was explored, but none yielded sufficient improvement, either in the log-likelihood or the interactions, to justify the computational cost associated. In these tests, batch sizes ranged from 100 to 5,000, hidden layer sizes from 36 to 76, values of k_{CD} from 1 to 11, and learning rates from 0.05 to 0.001. The log-likelihood and free energy of the trained machines are shown in Figure 2.12, and it can be seen that both the log-likelihood and the free energy indeed stabilised near the end of training, but that the log-likelihood decayed throughout training. These results were consistent across the different values for the hyperparameters.

The mean and standard deviation of the 1-, 2-, and 3-point interactions are shown in Figure 2.14. It can be seen that within statistics, the linear terms were indeed zero, though with large variance and a slightly negative bias. Similarly, the non-nearest neighbour 2-point interactions were zero within statistics. The nearest neighbour two-point interactions incorrectly showed a slightly negative coupling. This coupling was weak, however, compared to the 3-point couplings among the interacting triplets, which were accurately reproduced across the range of temperatures. Figure 2.13 shows the mean value of each kind of interaction during training. It can be seen that the machine started by fitting the data with linear terms, which then decreased and were replaced by pairwise nearest-neighbour interactions, which in turn decayed to be replaced by the correct 3-point interactions.

Epochs	Batch size	LR	k_{CD}	n_{vis}	n_{hid}
25k	200	0.05	1	64	64

Table 2.4: Training parameters for toroidal 8×8 Ising lattice with triplet coupling only, at $T = 1.5$ up to $T = 2.5$.

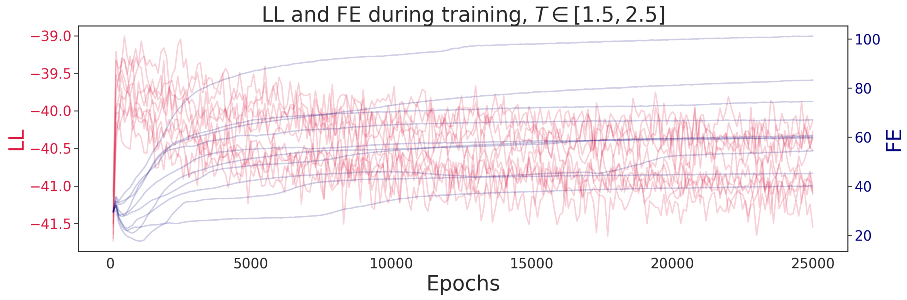


Figure 2.12: The free energy and log-likelihood across all 25k training epochs, for machines trained on Ising data with triplet interactions at a temperature ranging from 1.5 to 2.5. While the free energy stabilised towards the end of training, the log-likelihood decayed throughout training.

While the mean value of the 3-point interactions was correct, the precise structure was often incomplete. The left panel of Figure 2.15 shows that a single machine often missed some three point interactions. I trained 10 machines with the same hyperparameters on the $T = 2.5$ data set, and while most machines had missing 3-point interactions, taking

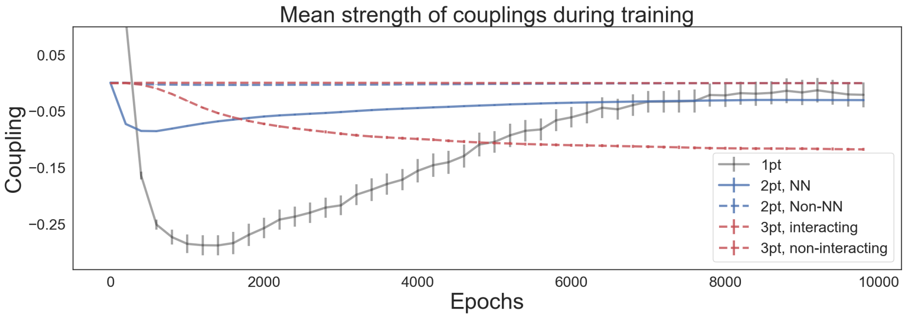


Figure 2.13: Couplings extracted from the machines across the first 10k training epochs. Errors bars are the standard error on the mean across the interactions, but only visible for the linear term. The machines encode the structure in the data in interactions at increasingly high order as training progresses.

the mean across these 10 machine recovered the ground truth structure and value. The mean estimated value of the triplet interaction $J^{(3)}$ was $-0.1313(\pm 0.0098, \text{sem})$, which includes the ground truth $\frac{J^{(3)}}{3T} = -0.133$.

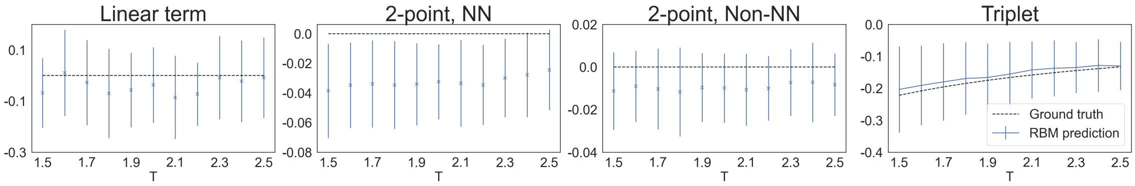


Figure 2.14: Couplings extracted from the machines at different temperatures. The encoded triplet interactions should have a coupling strength given by $J_{ijk} = \frac{J^{(3)}}{3T}$. Error bars correspond to standard deviations, and show that the ground truth was accurately reproduced for 1- and 3-point interactions, but that nearest-neighbours coupled slightly negatively, even when the ground truth was non-interacting.

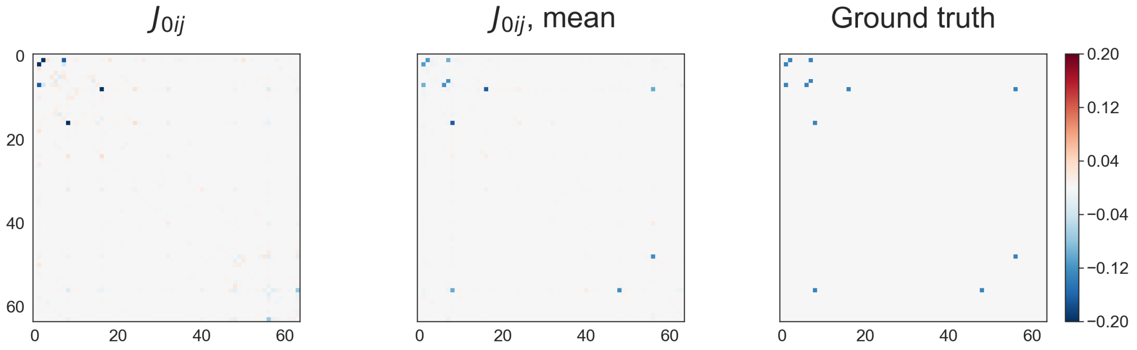


Figure 2.15: Slices of the three-point interactions J_{0ij} at $T = 2.5$ show that while a single machine often contained false negatives in its 3-point interactions, the mean across all machines accurately reproduced all triplet interactions. **Left:** Interactions from a single machine. **Middle:** Mean interactions across 10 machines. **Right:** Ground truth.

2.3.1.d Conclusion

I was able to reproduce the results from [59], and quantified the training using the log-likelihood, KL-divergence, and reproducibility across machines. Moreover, I have shown that RBMs can learn higher-order interactions, though they are prone to false negatives. This could be mediated by training multiple machines, and looking only at mean interactions. It is worth noting that the machines could still learn, even with a decaying log-likelihood. This might indicate that the log-likelihood becomes harder to estimate as weights increase. Finally, note that the RBM weights had a much larger spread than the extracted couplings. In fact, there did not seem to be any pattern that stood out among the weights. This relates to the more general issue of interpretability of weights in neural networks. Given that there is so much less degeneracy in the couplings, they could offer a more meaningful insight into the workings of neural networks than a direct analysis of the network weights.

2.3.2 Interactions in traits from the UK Biobank

2.3.2.a The RBMs trained to criticality and identified two phases

Training was divided in three eras of which the hyperparameters are reported in Table 2.6. The log-likelihood and free energy are shown in Figure 2.16. The log-likelihood initially increased, but then slightly decayed while the free energy stabilised. Figure 2.17 shows the distribution of the magnetisation of the training data, *i.e.* the mean value across traits. Comparing this with the magnetisation of 10^6 samples from the RBM, I noticed that I was able to reproduce these peaks, but in different ratios. Figure 2.18 shows why this was the case: the Gibbs sampling never thermalised, and the generated samples alternated between a phase with high magnetisation, and one with low magnetisation. Moreover, the autocorrelation of this time series did not decay to zero. It was therefore not possible to draw *i.i.d.* samples from the machine to compare its moments with those in the training data. This is indicative of a system that is at or near a critical point of a first order phase transition and in that context is called *critical slowing down*. A further sign of criticality was found in the distribution of generated states. I ordered all states by their probability under the RBM's model, and plotted their rank in this ordered list against the assigned probability. Figure 2.19 shows that the states followed a power-law distribution. The system did not have the pure Zipf exponent of 1, and especially in the tail of low probabilities the power law behaviour disappeared, but for the first 10^5 states the power law seemed to have a stable exponent of around 1.9.

Figure 2.20 shows the distribution of sample magnetisation during training, and it can be seen that the RBM learned the structure and position of these two peaks early on in

Total		Era 1			Era 2			Era 3		
Epochs	Batch size	Epochs	LR	k_{CD}	Epochs	LR	k_{CD}	Epochs	LR	k_{CD}
30000	5000	15000	0.01	10	6400	0.005	15	8600	0.001	20

Table 2.5: Training was separated into three eras.

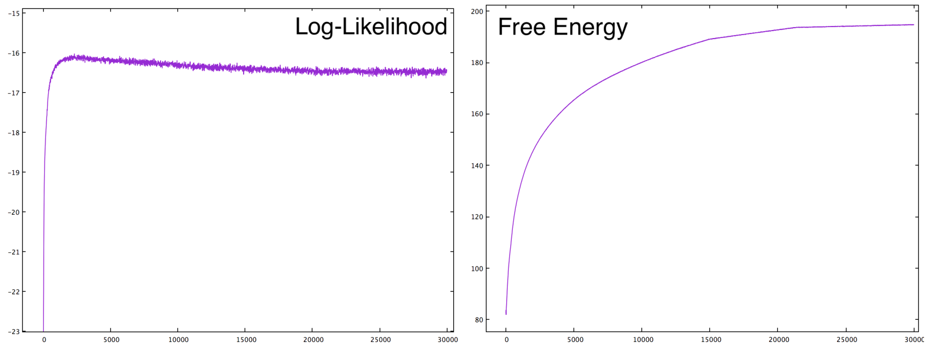


Figure 2.16: Training metrics for the UK Biobank machine. Note that the log-likelihood stopped growing very quickly, but the free energy stabilised only in the last era. (Figures provided by Ava Khamseh)

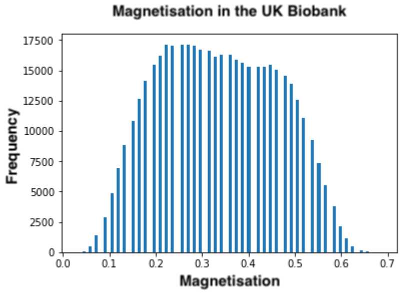


Figure 2.17: The histogram of magnetisations for the real UK Biobank data showed hints of multimodality.

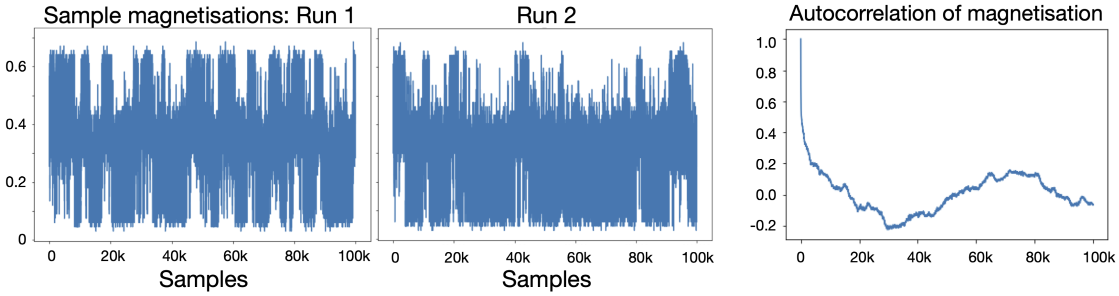


Figure 2.18: Two Gibbs sampling runs on a given trained RBM show that the consecutive samples were not independent, as confirmed by the autocorrelation on the right.

training, after around 100 epochs, or at 1% of training. To investigate these two phases further, I calculated the correlation between each trait and the full magnetisation across 10^6 Gibbs samples. Figure 2.21 shows that the two phases separated the samples into individuals with a high fat percentage throughout their body, low metabolic rate, that are female, and another group that are heavier, with less fat, a high metabolic rate, that tend to be male. Figure 2.22 shows that the traits that correlated strongly with magnetisation indeed divided the simulated cohort into the two peaks of the distribution, and that basal metabolic rate did so most cleanly. Indeed, when fixing the basal metabolic rate to 1 and Gibbs sampling from the conditional RBM distribution, the autocorrelation function

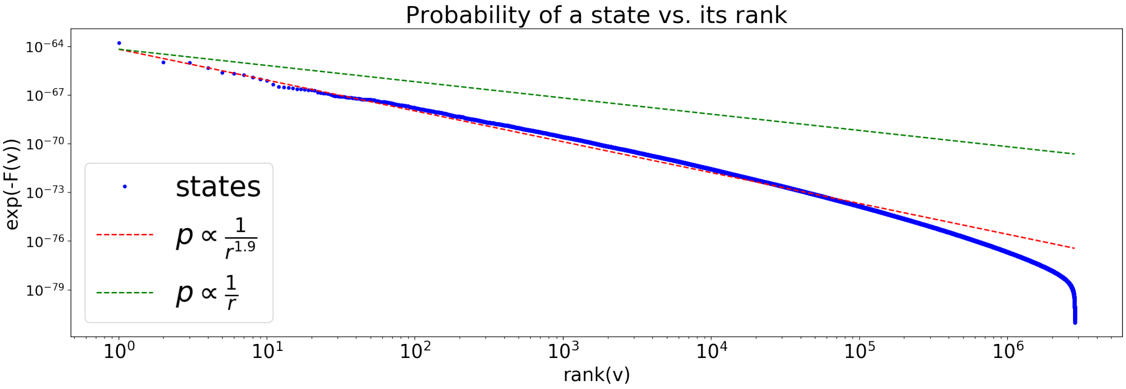


Figure 2.19: The Gibbs samples obeyed power law statistics with an exponent of around 2.0 (mean across 20 machines, only one shown) when ranking states by their probability under the RBM's model.

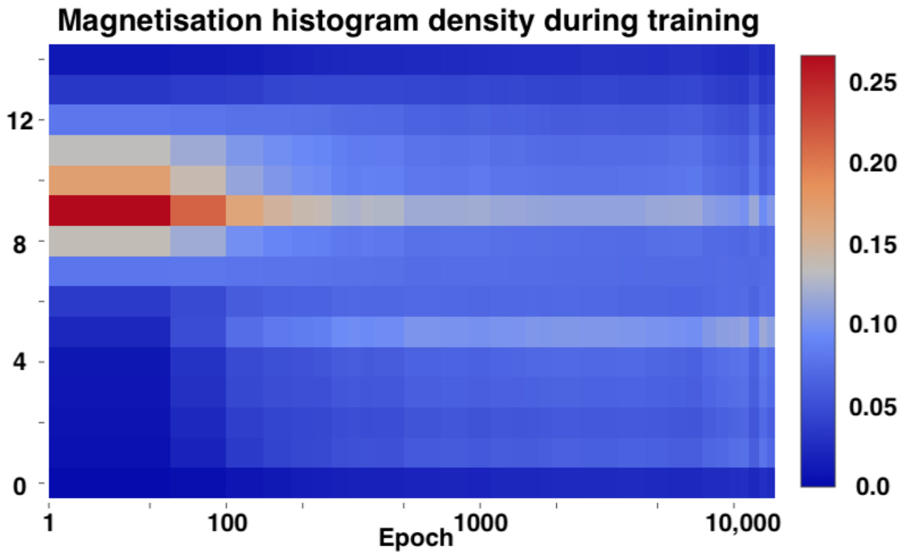


Figure 2.20: A heatmap of the histogram density during training. Note the logarithmic x-axis. The double peak appeared early in training, after around 100 epochs.

decayed to zero within 200 samples.

In spite of the usual training metrics not being effective, the encoded couplings could still be meaningful. The linear terms, or 1-point interactions, are shown in Figure 2.23. It can be seen that most were positive, but more importantly, that they were mostly consistent across 20 identically trained machines. The 2-point couplings with *basal metabolic rate* (BMR) are shown in Figure 2.24. A similar pattern was observed in other couplings: many couplings were spread around zero, but some showed similar nonzero couplings across the 20 machines. Training on a data set where all traits were randomly shuffled to destroy all correlations gave couplings of $\mathcal{O}(10^{-3})$, which could serve as a background against which many of the couplings are significantly nonzero. The strongest interactions all seemed to coincide with biological intuition. For instance, high BMR coupled to, and is epidemiologically associated with, lower fat percentages and higher lean mass. Still, there were some surprising negatives: BMR did not couple to sex or grip strength. Taking the mean across the 20 machines, the full two-point coupling matrix is shown in Figure

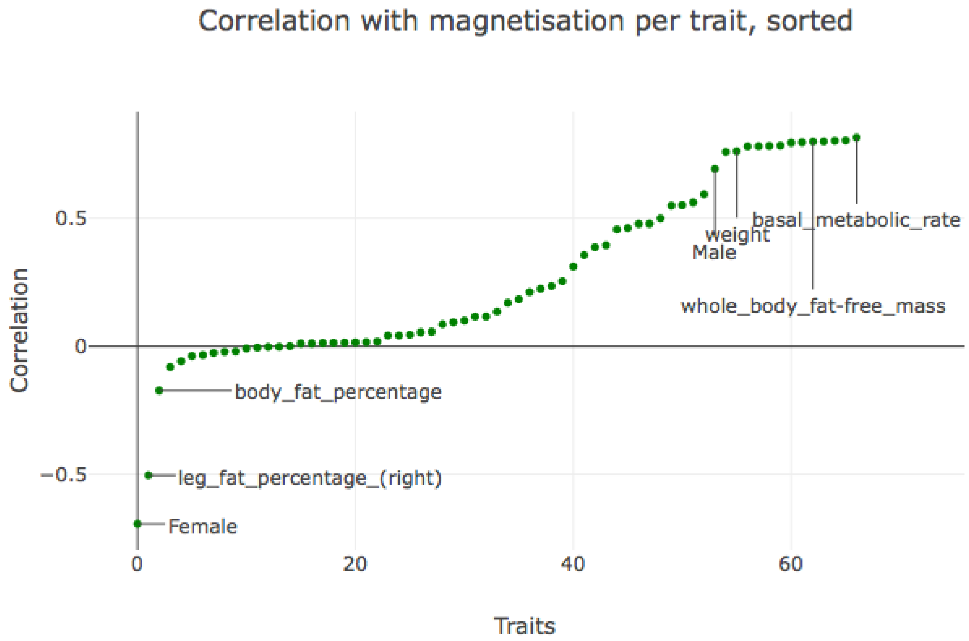


Figure 2.21: The correlation across 10^6 Gibbs samples between magnetisation and each of the traits. The traits that were negatively correlated with magnetisation were all female-associated, while high correlation coincided with male-associated traits.

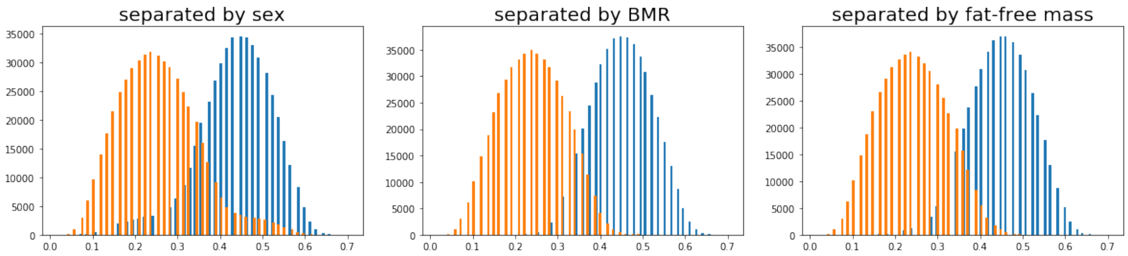


Figure 2.22: Magnetisation histograms of a simulated cohort, separated by different traits. While the modes seemed to correspond to sex-associated traits, BMR separated the two peaks most cleanly.
Left: Male (blue) and Female (orange), note the longer tails in each distribution.
Middle: High BMR (blue) and low BMR (orange)
Right: High whole body FF-mass (blue) and low whole body FF-mass (orange).

2.25. The RBMs captured many obvious relationships: the mutually exclusive age and sex traits coupled strongly, and negatively. The control-traits (skin colour, number of vehicles, and tea-intake) did not couple to any of the other traits. The different traits related to blood-counts formed a set that mostly couple among themselves, as did the fat- and mass-associated traits.

Finally, the 3-point interactions between the traits *female*, *BMR*, and any other trait are shown in Figure 2.26. There were a few traits that showed a clear signal across the 20 machines, mostly relating to fat-free mass, reflecting a relationship between sex, metabolic rate and muscle-mass that is more complex than the sum of pairwise interactions.

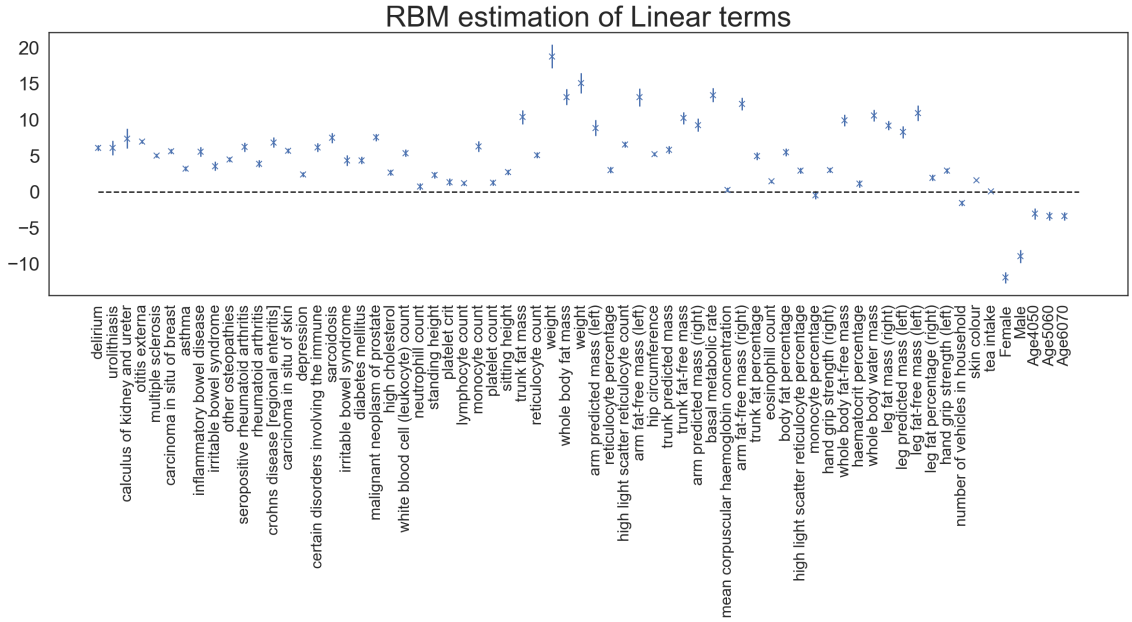


Figure 2.23: Linear term for each trait. Error bars denote standard deviation across the 20 machines.

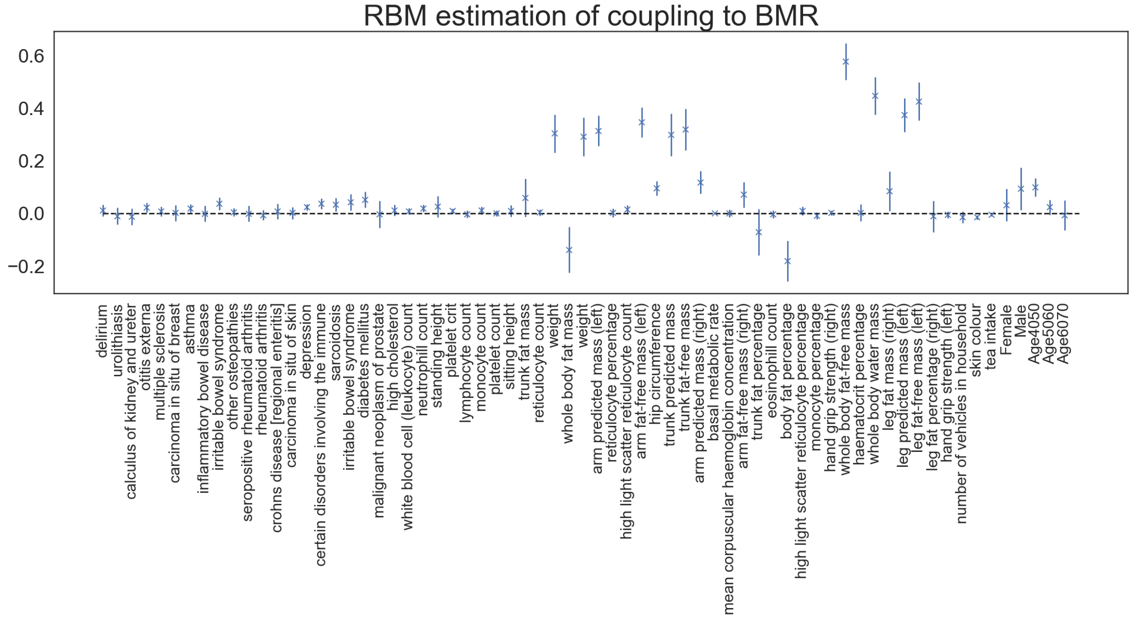


Figure 2.24: Couplings with BMR for each trait. Error bars denote standard deviation across the 20 machines.

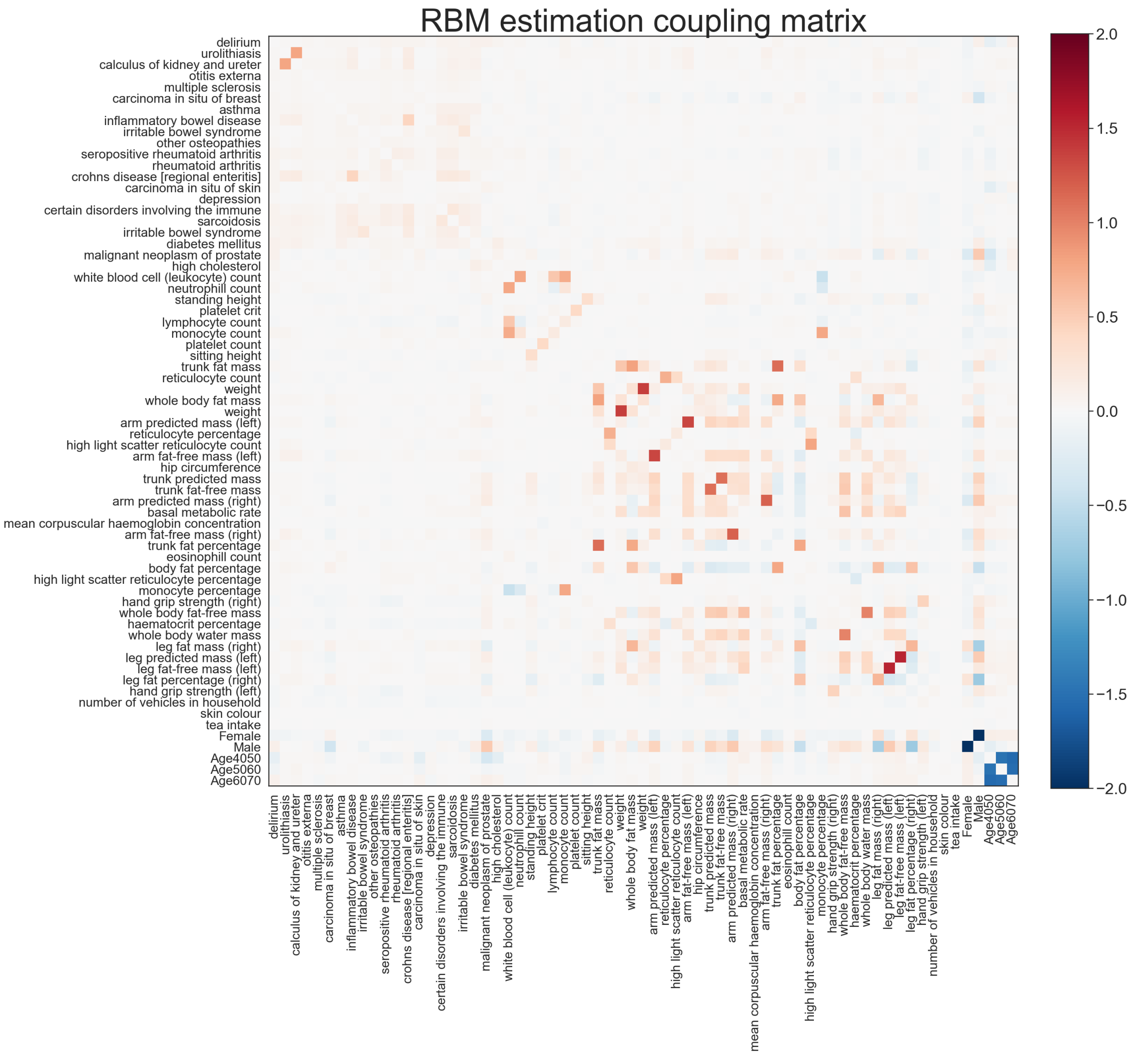


Figure 2.25: Two-point coupling matrix across all 67 traits, mean across 20 machines.

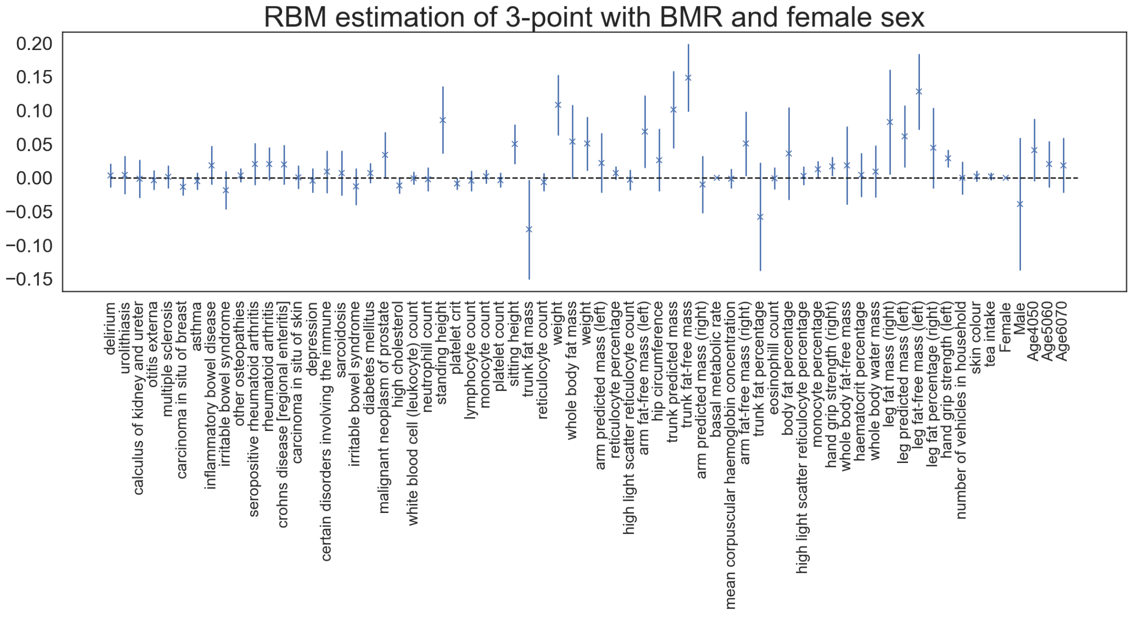


Figure 2.26: Three point couplings with BMR and Female.

2.3.2.b 2-point interactions found a link between Crohn's/IBD and kidney stones

In many situations, the strength and sign of the 2-point couplings can be hypothesised to agree with the strength and sign of the pairwise correlation. However, since correlations are different from the couplings in various respects, the pairs of traits that do not follow this trend are of particular interest. Under the assumption that there is a linear relationship between coupling and correlation, the outlier points—those with a higher or lower value of coupling than expected—should reveal novel dependencies in the data not accessible by pairwise correlation. To find these, I fitted a linear model for each trait to predict the couplings from the correlations and defined outliers by their Cook's distance. The Cook's distance combines the idea of having a large residual and a high influence. It is defined as

$$C_j = \frac{\sum_{i=0}^{n} \left(\widehat{y}_i^{(j)} - \widehat{y}_i \right)^2}{ps^2} \tag{2.31}$$

Where $\widehat{y}_i$ is the prediction for data point i of the full linear model, and $\widehat{y}_i^{(j)}$ is the prediction for that point if point j had been removed prior to fitting. The total number of observations is n, p is the number of parameters in the linear model ($p = 2$ in this case) and s^2 is the mean squared error of the original fit. The outliers can then be defined as having a Cook's distance larger than some threshold. An oft-mentioned rule of thumb is that points should be considered influential if $C_j > 4/n$.

The data is binary, so an appropriate measure of correlation has to be chosen. I considered the following: Pearson correlation, Spearman correlation, Covariance, Tetrachoric correlation, Matthew's correlation, Sokal-Sneath distance, Sokal-Michener distance, Hamming distance, and Jaccard distance.

I set the C_j threshold at $4/67 = 0.06$, and the correlation *vs.* coupling plot for *Calculus of kidney and ureter* is shown in Figure 2.27. While different measures of correlation led to different plots and fits, in all but one case, and across thresholds from $[0.03, 1.2]$, the same two nontrivial outliers were identified. *Urolithiasis* was always an outlier, which makes sense considering the fact that it is a superset of *Calculus of kidney and ureter*. However, the coupling between *Calculus of kidney and ureter* and *Crohn's disease* and *IBD* was not reflected in most correlation measures.

2.3.2.c 2-point couplings removed sex-bias and reveal cancer comorbidities

Figure 2.28 shows outliers in the Pearson correlation vs. 2-point coupling plots for four traits. Some outliers were not very informative: It is no surprise that *Male* was positively associated to prostate cancer, and negatively associated to breast cancer. However, note that it was much clearer from the couplings than from the correlations that males are more susceptible to diabetes mellitus. Prostate cancer was negatively correlated with skin cancer (perhaps because of an age- or survival-related confounder), but was its strongest positive non-age-related interactor. It has indeed been observed that melanoma is linked to prostate cancer [187]. Similarly, BMR and breast cancer were negatively correlated (probably because men tend to have higher BMR but lower incidence of breast cancer),

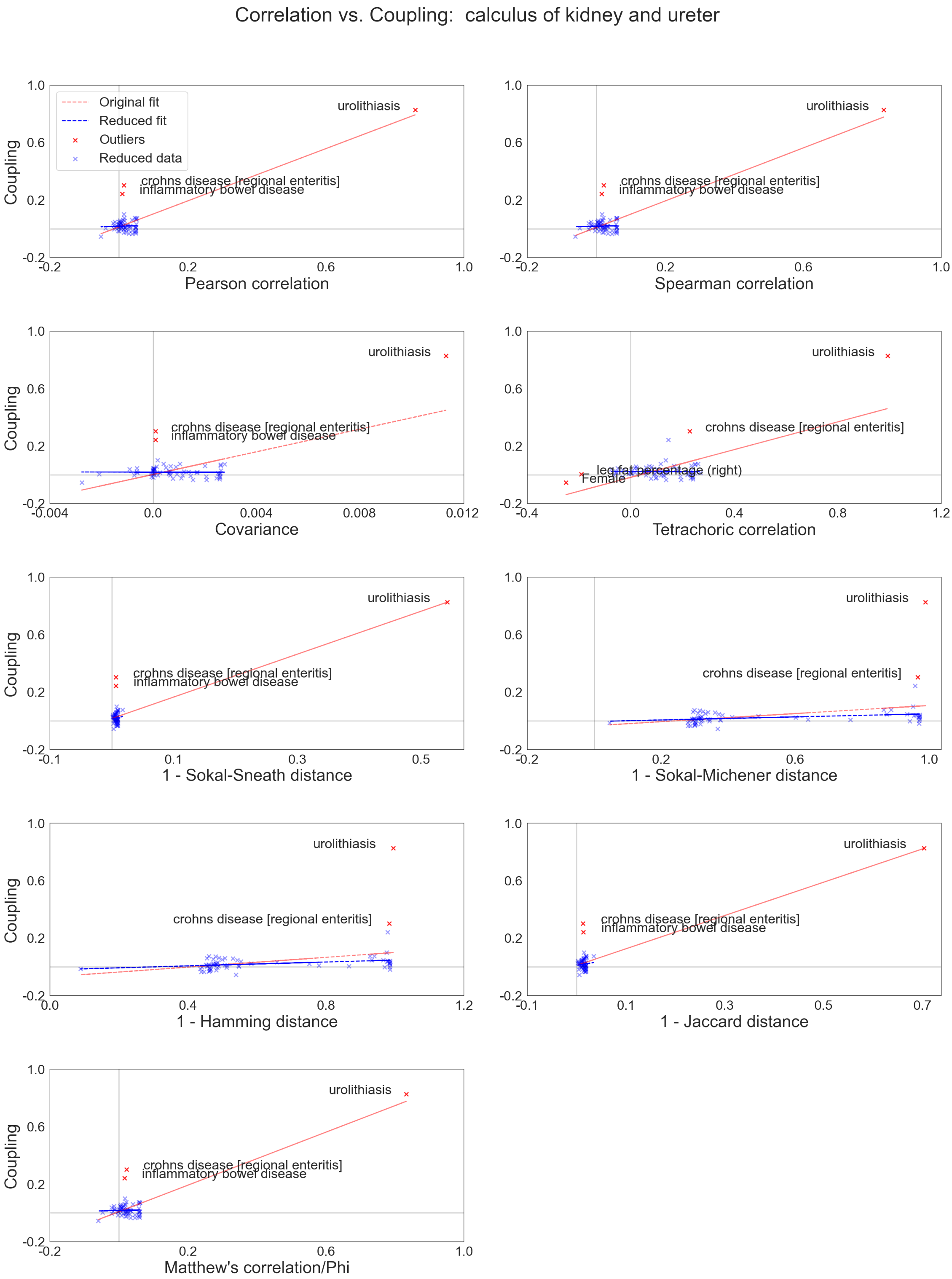


Figure 2.27: Outlier plots based on Cook's distance for 'Calculus of kidney and ureter' with different distance measures. Some traits—such as *Calculus of kidney and ureter*, *Crohn's disease*, and *IBD*—are consistent outliers across various distance measures, which means that their strong coupling was not reflected by a correlation.

but coupled positively. It has indeed been observed that BMR is positively associated with risk of breast cancer [136].

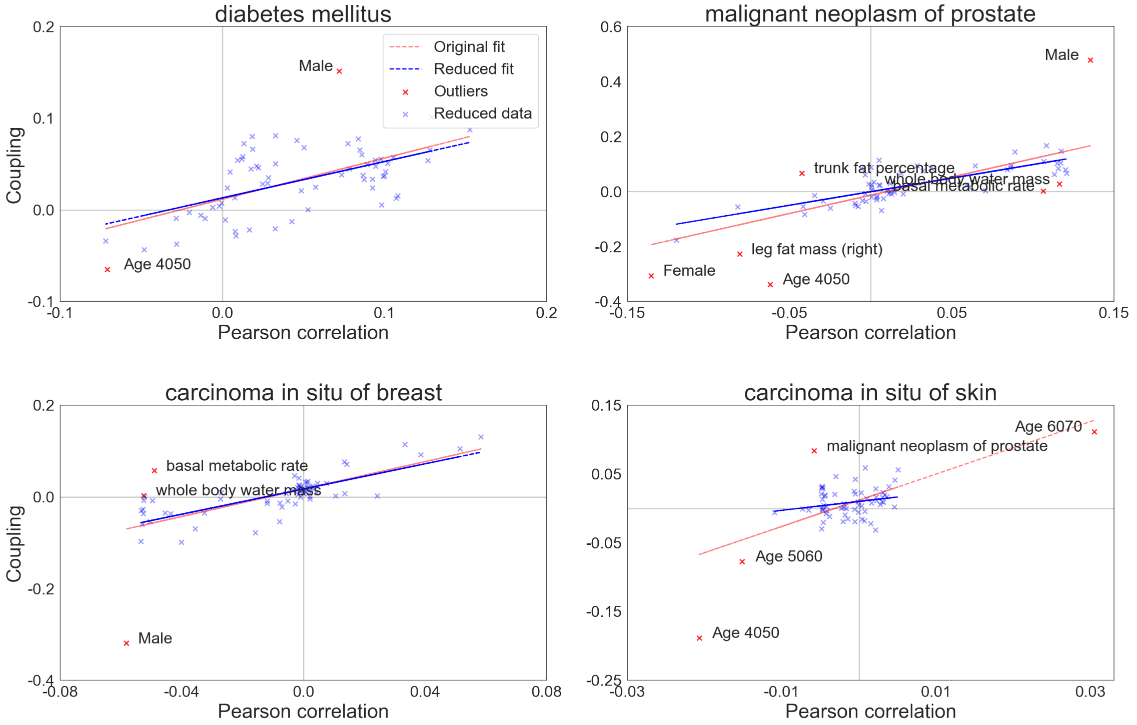


Figure 2.28: The correlation vs. coupling plots for four different traits shows how outliers affect the regression line. The red line is the fit on all data, the blue line is the fit once the outliers (red points) have been removed. The C_j threshold was set at $4/67 = 0.06$.

2.3.2.d Conclusion

To conclude, I found significant interactions among the 67 traits that showed biological plausibility. To reproduce the training data, the RBMs had to encode a model with critical dynamics in the visible layer, hindering the generation of *i.i.d.* samples. These critical dynamics were necessary because the training data comprised two modes, most likely corresponding to the two sexes (although BMR seems to separate the two phases better). I found instances of pairs of traits that were negatively correlated, but coupled positively and are indeed positively associated in the literature. This seems to imply that the learnt interactions were accurately controlling for confounding variables in the data set.

2.3.3 Interactions in developmental astrocytes

This section describes the interactions encoded by RBMs trained on gene expression data from late developmental astrocytes. As shown in the previous sections, the log-likelihood is not necessarily a good training metric, but when training on gene expression data, there is no ground truth to compare with. As an alternative measure of accuracy and robustness, I considered different sources of variability in the couplings:

- Machines: Machines that only differ in their weight initialisation and the stochasticity of gradient descent should learn the same couplings, since a Boltzmann distribution is completely determined by the couplings.
- Cells: Machines that are trained on different, but biologically equivalent, subsets

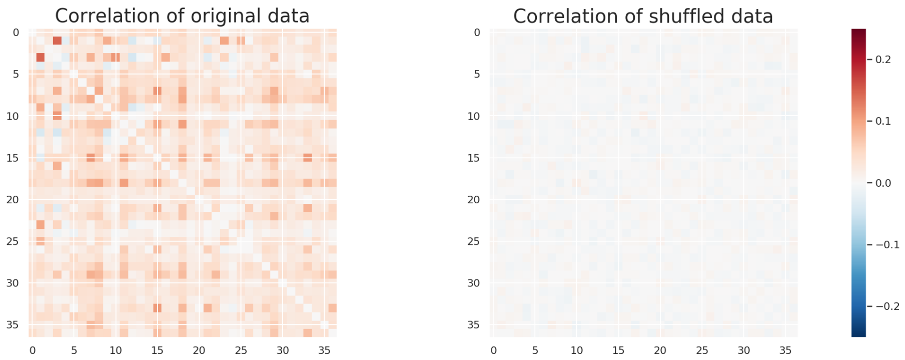


Figure 2.29: A comparison of gene-gene correlations in the original and shuffled data sets shows that the shuffling correctly destroyed all correlations.

of the cells should learn the same couplings if the couplings are representative of true biological mechanism.

- Genes: Machines that are trained on different subsets of genes should learn the same couplings between the genes in the intersection.

To be able to address these sources of variability, I trained 20 identical machines on each of eight different data sets:

1. **CL8:** The full data set of all 77,524 10X astrocytes (cluster 8 of the k=20-means clustering), with the 37 highly expressed, astrogliogenesis-associated genes.
2. **DS1:** A random selection of half of these cells (38,763 cells in total).
3. **DS2:** The other half.
4. **Shuffled:** The full data set, but with each gene's expression shuffled across cells to destroy all gene-gene correlations (see Figure 2.29).
5. **nG5:** All cells, but only the first 5 genes.
6. **nG15:** All cells, but only the first 15 genes.
7. **nG25:** All cells, but only the first 25 genes.
8. **3pt:** This data set is only used in Appendix 2.B. It contains all cells, the 37 genes, and 5 additional simulated variables. To generate values for these 5 additional variables, I sampled from an Ising model with three-point couplings (as was done in Section 2.3.1). The first 3 of these 5 simulated genes couple with a three point coupling, while the last two do not couple to anything. This should reveal a machine's ability to recognise and estimate three-point couplings.

2.3.3.a Training scheme and learning accuracy

I trained 20 machines on each of the data sets. Each machine had $n_{\text{vis}} = n_{\text{hid}}$, and was trained with contrastive divergence according to the scheme in Table 2.6. The hyperparameters were chosen after a gridsearch on the CL8 data set with $k_{CD} \in \{1, 5, 10\}$, $n_{hid} \in \{30, 37, 60\}$, batch size $\in \{5, 20, 50, 100, 200, 500\}$, and learning rate

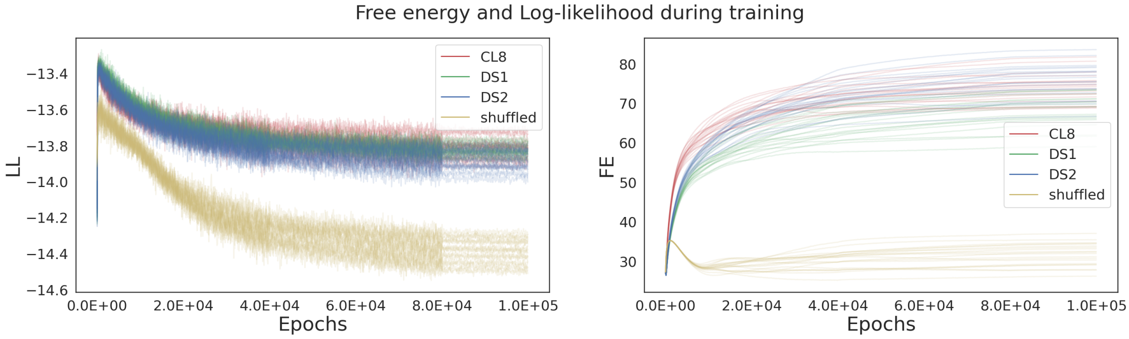


Figure 2.30: The log-likelihood and free energy of RBMs trained on the various gene expression data sets. As before, the free energy stabilised at the end of training, but the log-likelihood decayed throughout.

$\in \{0.05, 0.01, 0.001\}$, using the first two moments of the training data as the evaluation criterion.

Total		Era 1			Era 2			Era 3		
Epochs	Batch size	Epochs	LR	k_{CD}	Epochs	LR	k_{CD}	Epochs	LR	k_{CD}
100k	200	40k	0.01	5	40k	0.005	5	20k	0.001	10

Table 2.6: Training scheme for the machines trained on the gene expression data sets.

For each of the 20 machines, the log-likelihood and the free energy during training are shown in Figure 2.30. Similar to the machines trained on simulated Ising data with triplet interactions (Section 2.3.1), the free energy stabilised as training progressed, but the log-likelihood decayed. Again, this either means the machines did not correctly train, or that the log-likelihood was not a useful training metric in this case. To evaluate how well these machine captured the training distribution, Figure 2.31 shows a comparison between the moments of the training data and the moments of 100k samples from the RBM (keeping only every 20th sample to eliminate the autocorrelation). Most machines were able to reproduce up to six moments of the training data.

Finally, while the stable log-likelihood and free energy suggest that the training procedure reached a stable point, the couplings should also reach a stable value. Figure 2.33 shows that both the 2- and 3-point couplings reached stable values after an initial period of large changes. Figure 2.34 shows the distribution of these changes across all couplings during training, which confirms that the changes were strongest at the beginning, and then stabilised. Note that this was most apparent for the true data set (CL8), and less apparent for the shuffled data set. Figure 2.35 shows the mean of the absolute value of the couplings, order by order, and how these changed during training. The first thing to notice is that throughout training, the shuffled data had weaker interactions than the original data set. Furthermore, the same relay-race pattern that was seen in the Ising interactions (*cf.* Figure 2.13) was present here. In the original data set the linear terms reached a maximum after around 10k epochs, while the 2-

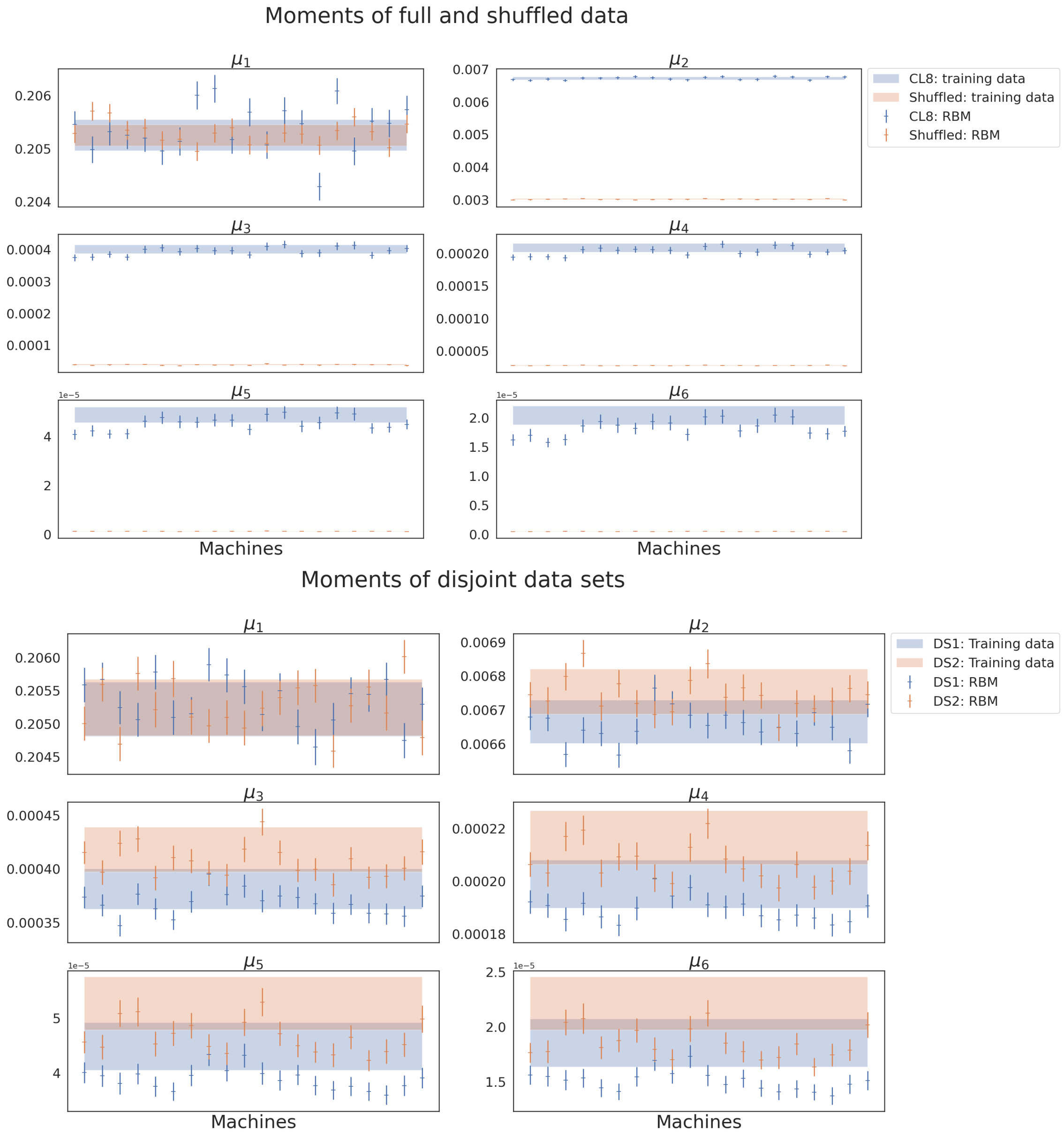


Figure 2.31: First six moments of the training data and trained machines from the data sets with the original, split, and shuffled data. The shaded region and the error bars cover the moments $\pm$ one standard deviation across 1000 bootstrap resamples of the training data and 1000 samples from the RBM, respectively.

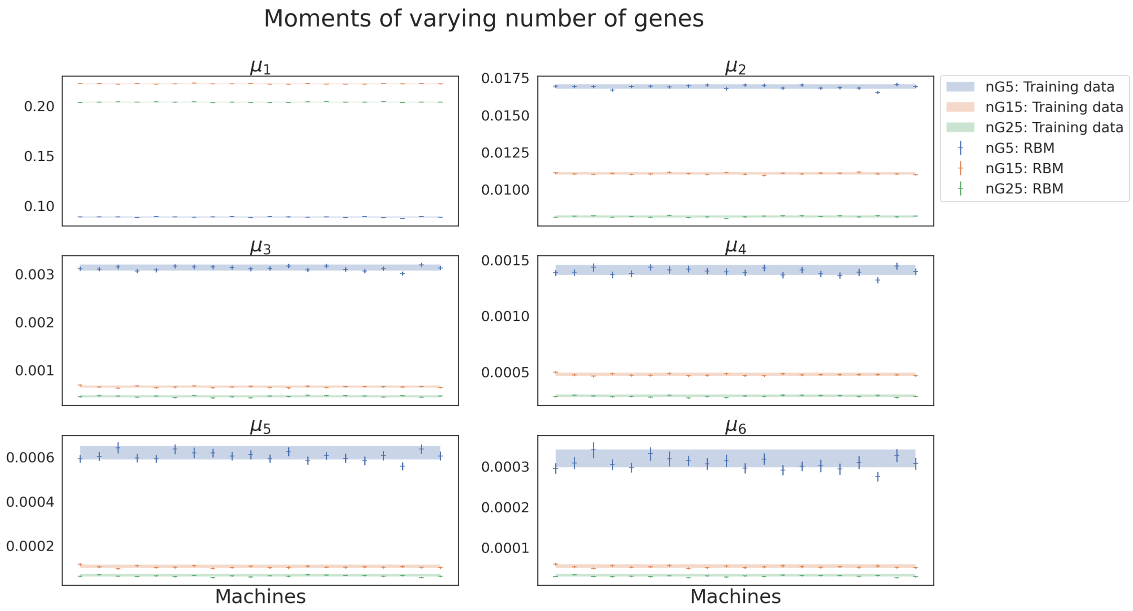


Figure 2.32: First six moments of the training data and machines from the data sets with a varying number of genes.

and 3-point interactions were still growing. Then, the linear terms started decreasing in absolute value, while the 2-point interactions reached a maximum at around 20k epochs. After this, the 2-point interactions started decreasing, while the 3-point interactions kept growing until they stabilised at their maximum value. This relay-race between interactions at different orders was completely absent when training on shuffled data. There, the orders did start growing at different moments, but none decreased to make room for others. This indicated that there was structure in the CL8 data set that the RBMs captured with increasingly complex models. The fact that the 3-point interactions did not decrease anymore was also an indication that it might be the highest order of interaction that is commonly present in this data set. This model of increasingly sophisticated understanding of the structure in the data might explain why this pattern was absent in the shuffled data, where there was no such structure to learn. It was further corroborated by the pattern in Figure 2.36, which shows that as training progressed, the machines reproduced higher moments of the training data.

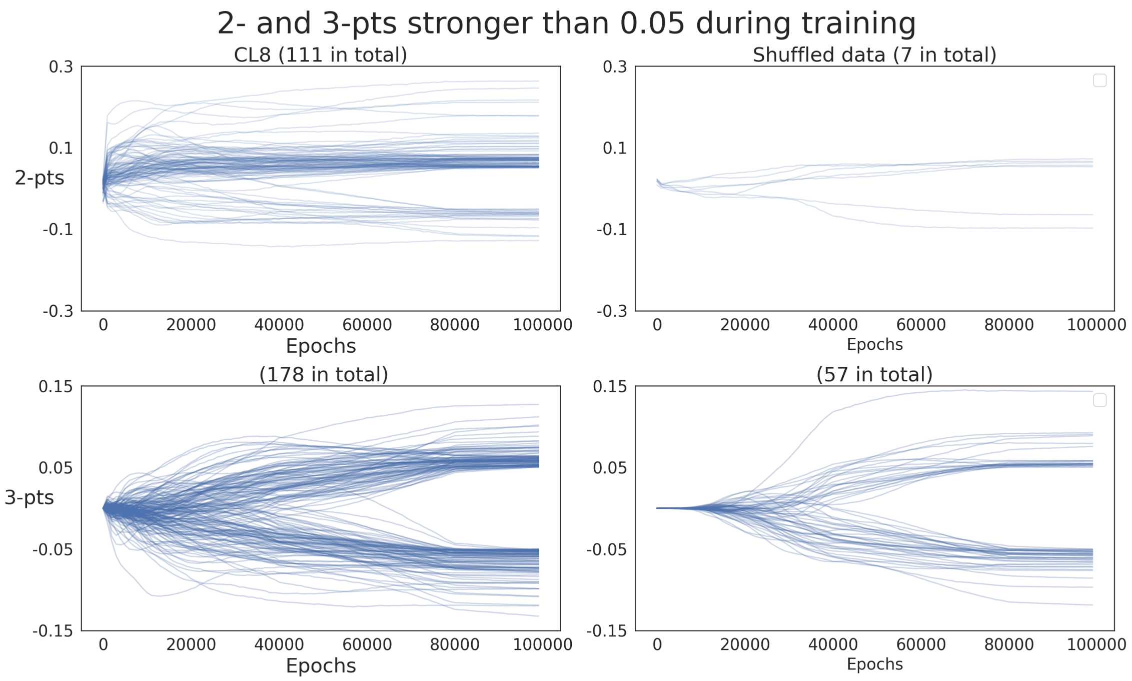


Figure 2.33: Values of individual interactions during training. Only shown for interactions that at the end of training had an absolute value above 0.05.

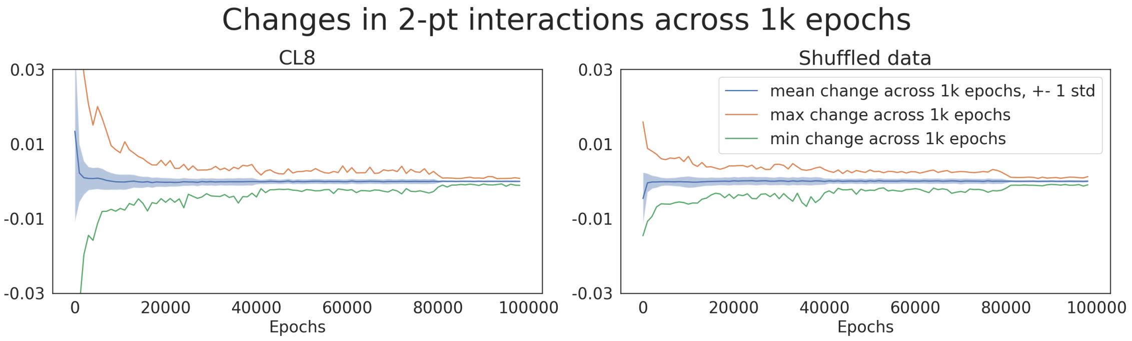


Figure 2.34: Changes in 2-point interactions during training.

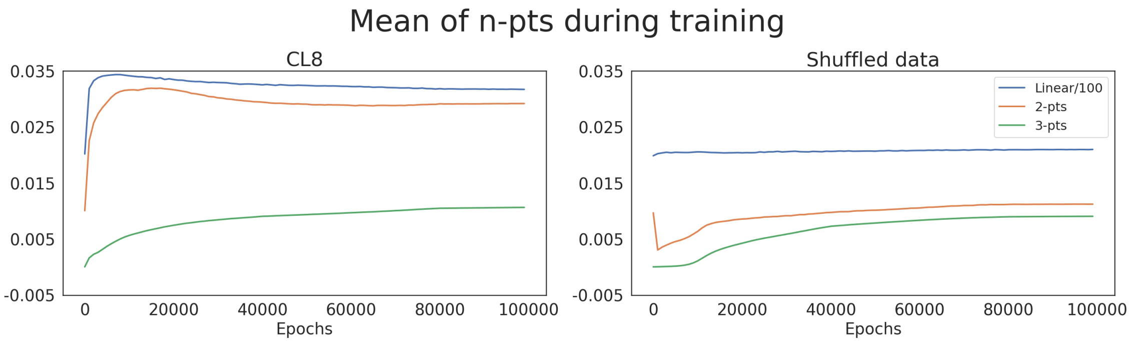


Figure 2.35: The different orders of interactions grew at different rates. Note that the linear term is divided by 100. In the original data set (CL8), the lower orders grew first, but got replaced by higher-order interactions later. In the shuffled data, there was no such relay-race present.

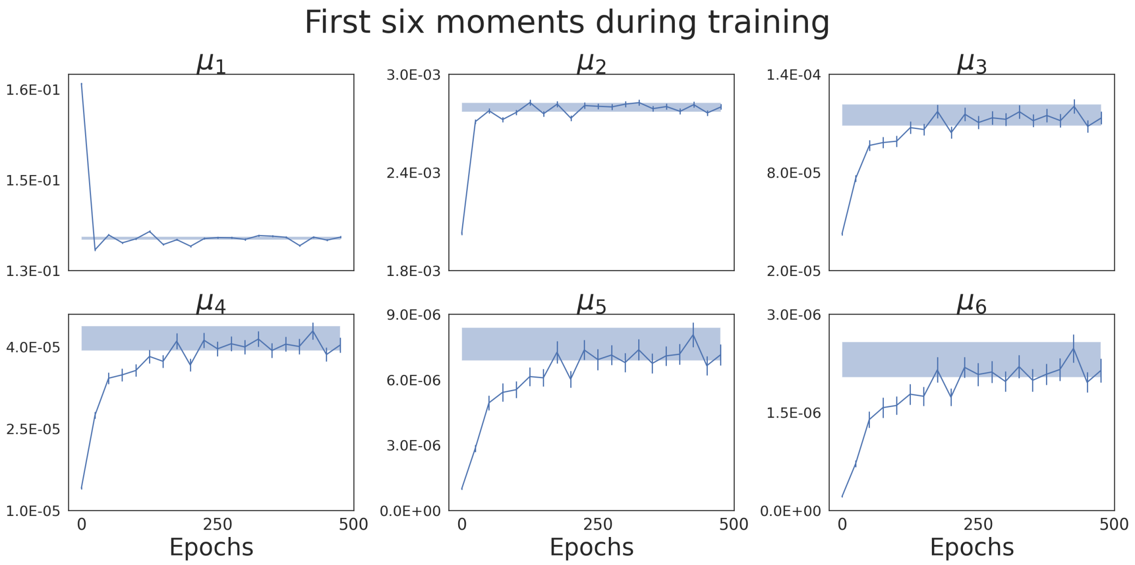


Figure 2.36: As training progressed, the machines learned increasingly higher-order structure of the training data. Shaded regions correspond to moments of the training data, lines to the mean across 20 machines. Error bars throughout denote ± one standard deviation. RBM moments were estimated on 1k Gibbs samples of the visible layer, thermalising the chain between each sample.

2.3.3.b Variability across machines

I first considered variability in couplings across identically trained machines. The 20 machines were trained with identical hyperparameters (see Table 2.6), on identical data. The only sources of variability were the initial weight configuration (sampled from $\mathcal{N}(\mu = 0, \sigma^2 = 0.01)$) and the stochasticity in batch gradient descent. Since the machines were trained on the same training distribution, and a Boltzmann distribution is uniquely determined by its couplings, the machines were expected to encode identical couplings. Note that the distribution is *not* uniquely determined by the weights of the RBM, as previously illustrated in Figure 2.10. Figure 2.37 explicitly shows the distribution of coupling across the 20 machines, once shown for all interactions with the weak interactor *Creb1*, and once with the strong interactor *Atf3*. While the variation across machines was significant, there were some couplings with a strong signal. In particular, the interactions between *Atf3* and the genes *Sned1, Jun, Ddit3, Gas7* and *Hsbp1* showed a strong positive signal. The interactions between *Atf3, Jun* and *Ddit3* were directly confirmed by the Pathway Commons database [219] as protein-protein interactions, and the String database includes at least low-confidence associations for each of the remaining interactions through homologues in other organisms. Still, the figures show that a single machine cannot indicate a significant interaction, as some couplings varied wildly, attaining large absolute values in some machines but averaging to around zero (e.g. *Atf3* and *Aqp4*). Therefore, I only considered the mean coupling across the trained machines as informative.

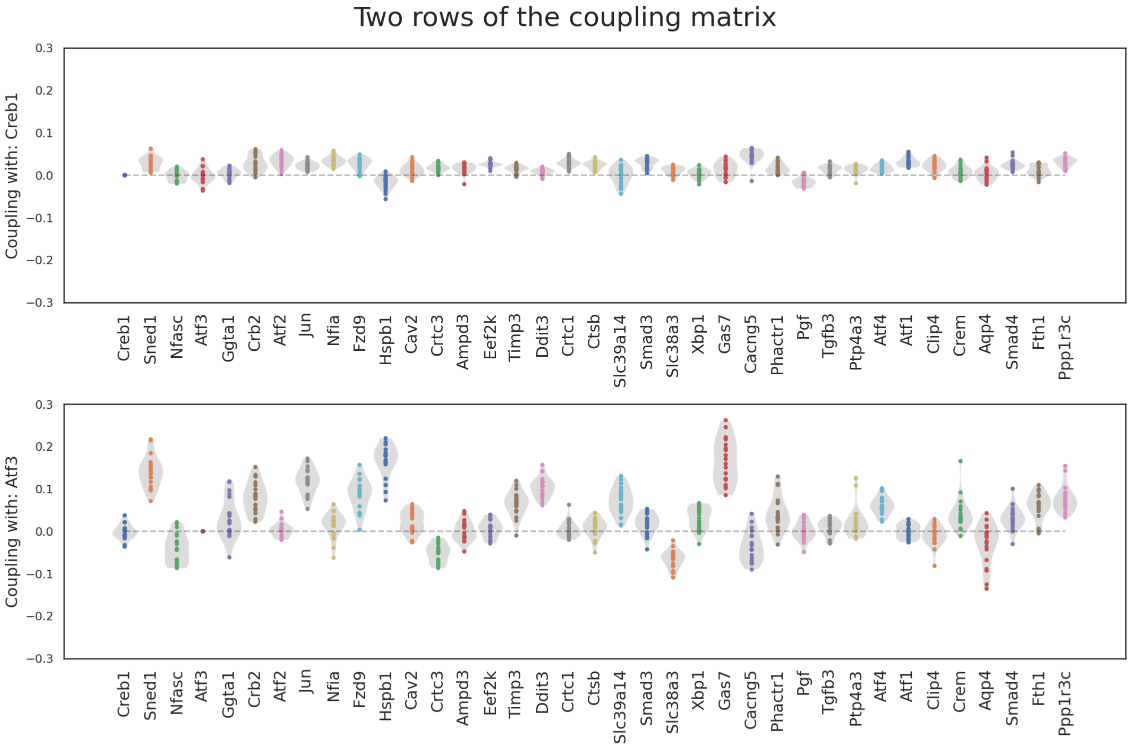


Figure 2.37: Two rows of the coupling matrix as violin plots across 20 machines. One row corresponding to a weakly coupling gene (*Creb1*), and one to a strongly coupling gene (*Atf3*).

2.3.3.c Variability across cells and genes

If the couplings are to be interpreted as reflecting true and direct dependencies between the genes, then they should be robust to a changing number of cells and genes in the training data. Figures 2.31 and 2.32 already showed that the RBMs were able to reproduce up to six moments of the training data, but here the robustness of the couplings themselves with respect to changes in the training set are investigated. Figure 2.38 shows a comparison of the coupling matrices of machines trained on different data sets. In line with the conclusion from the previous section, only mean values across the 20 machines are shown. Visually, the coupling matrices appear stable across the two disjoint data sets DS1 and DS2, and the data sets that only included a subset of the genes. In Table 2.7 this similarity is quantified by the normalised Frobenius norm of the difference, defined for two matrices M and N as $\frac{||M-N||}{||\frac{1}{2}(M+N)||}$, where $||M||$ is the usual Frobenius (or Euclidean) norm of a matrix. The numbers in Table 2.7 confirmed the similarities in Figure 2.38. Most dissimilar were the coupling matrices of the real CL8 data and the shuffled data. Next in terms of dissimilarity were DS1 and DS2. The interactions based on subsets of genes were all more similar to each other than DS1 and DS2, showing that the couplings are robust to training on these subsets.

Because the interactions seemed robust to training on subsets of genes, I also trained 20 RBMs with identical hyperparameters on a superset of genes—the union of the original 37 and the 100 most highly variable genes that are expressed in at least 2k cells. This created a new data set with 136 genes in total. This significantly changed the coupling matrix, and yielded a much worse Frobenius difference with the original data set, as

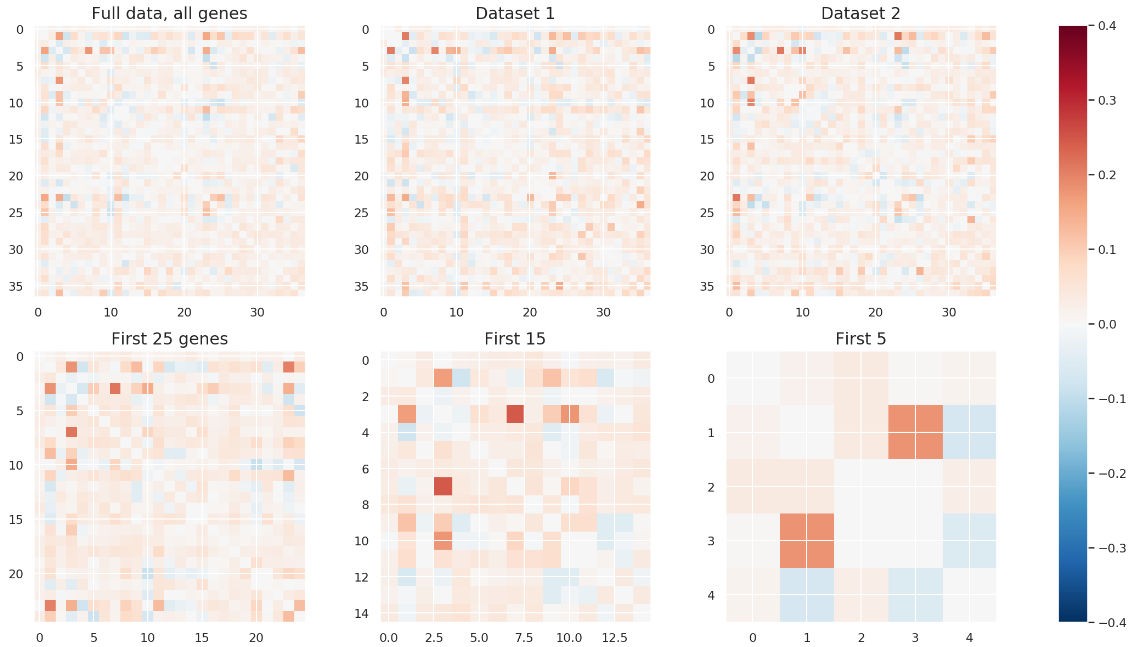


Figure 2.38: Mean couplings across 20 machines, for each of 6 data sets. The top row corresponds to machines trained on different cells, the bottom row to machines trained on different genes.

shown in the bottom row of Table 2.7. The coupling matrices in Figure 2.39 show that while some patterns persist, the magnitude of all couplings decreases, and most get lost in the background noise. I conclude that while the couplings were robust with respect to the subsets of genes, they are not robust to this larger superset.

Matrices	Frobenius difference
C_{37} − Shuffled	2.0879
DS1 − DS2	0.7426
C_{37} − DS1	0.5021
C_{37} − DS2	0.5332
$C_{37} - \frac{1}{2}(\text{DS1} + \text{DS2})$	0.3320
$C_{37} - C_5$	0.5677
$C_{37} - C_{15}$	0.6045
$C_{37} - C_{25}$	0.5929
$C_{37} - C_{136}$	1.2993

Table 2.7: Normalised Frobenius difference between matrices. The matrix C_i is the mean coupling matrix when training on i genes. Each difference is restricted to the intersection of selected genes.

This dependence on the specific genes selected means that there was no way to estimate the significance and robustness of any particular interaction, unless the RBMs are trained jointly on a much larger set of genes. It is indicative of a strong omitted variable bias.

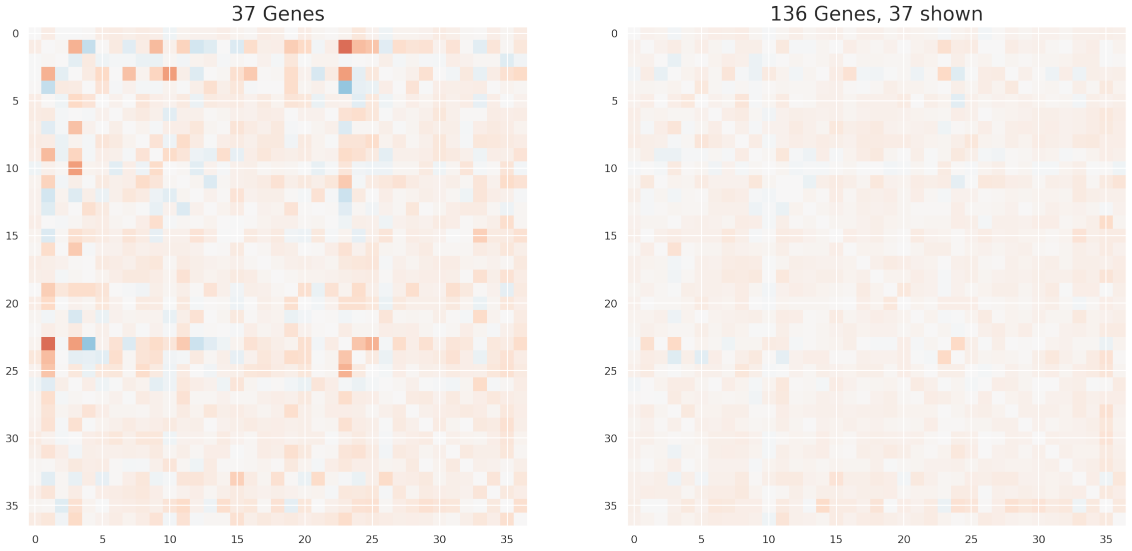


Figure 2.39: Original coupling matrix and the same couplings when training on a superset of 136 genes. Most structure is lost to background noise.

2.3.3.d Conclusion

I used RBMs to estimate genetic interactions at 1st, 2nd, and 3rd order from single-cell expression data. The log-likelihood was not a useful training method as it decayed even when the RBMs were getting better at reproducing the data, so other metrics had to be used to evaluate performance. The trained machines reproduced up to six moments of the data, and the interactions were robust to retraining on subsets of genes and cells. However, the interactions were not robust to retraining on a superset of genes. Because of this, the approach no longer seemed fruitful and I did not further analyse the 3-point interactions in this data, but in Appendix 2.B I show that when artificial 3-point interactions are added into the data set, the RBMs were able to reproduce their structure.

2.3.4 Parallel tempering does not prevent LL decay

On all data sets from this Chapter, except for the pairwise-only Ising models, the log-likelihood decayed during training. Two possible explanations for this are inaccurate estimation of the partition function, or inaccurate estimation of the gradients. The latter would also lead to increasingly bad estimates for the moments and interactions, but this was not observed, so is already deemed less likely. To eliminate the possibility that the gradients are not estimated accurately, I investigated the effect of improving the gradients with a method known as parallel tempering. To calculate the gradients, the CD algorithm approximates the expectation value over the model distribution by the mean across Gibbs samples, starting the sampling chain at one of the training examples. This is not an unbiased exploration of the model distribution, so could be the source of the errors in the gradient estimation. I retrained machines on the pairwise Ising models and the gene expression data, using parallel tempering (PT) to estimate the gradients, to see if it improved training convergence.

I created a new data set of 25k Ising states at $T = 1.8$, on an 8×8 toroidal lattice with

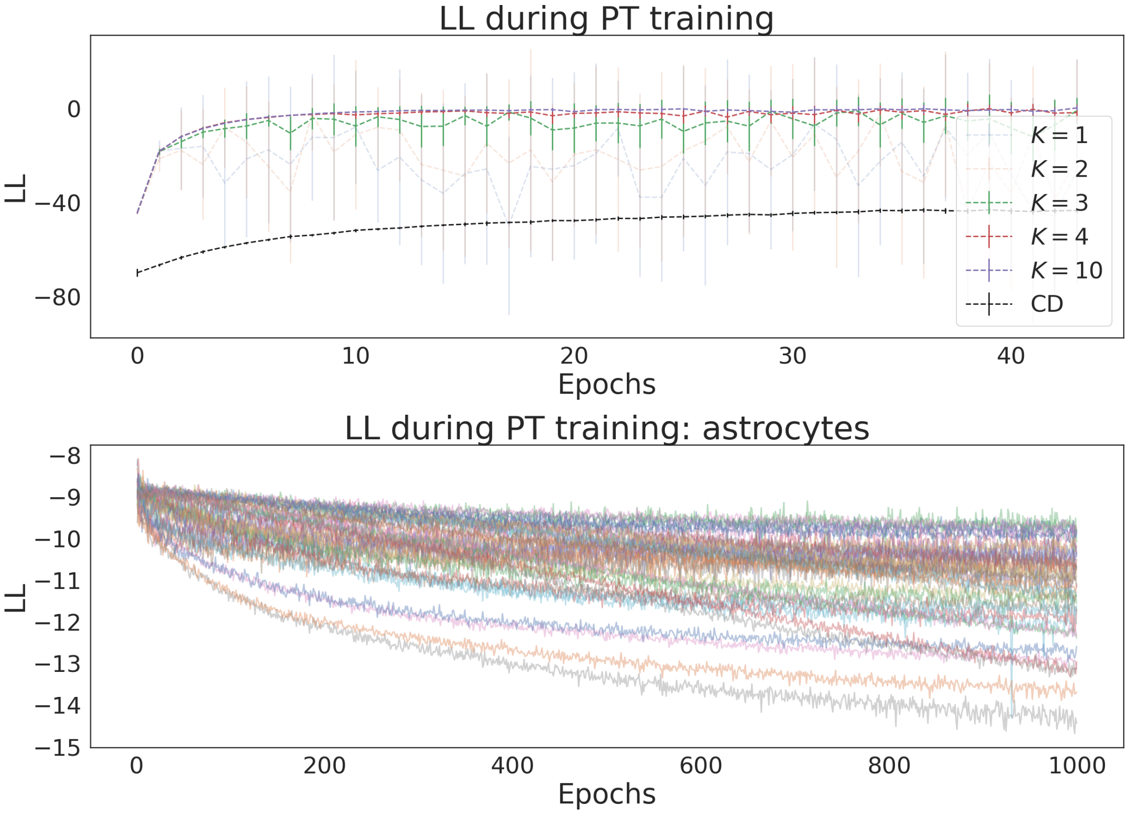


Figure 2.40: The log-likelihood during training, with error bars corresponding to standard deviations across 10 machines. Both figures show a comparison between contrastive divergence (CD) training, and PTK, i.e. parallel tempering with K parallel chains. The top figure shows the log-likelihoods of machines with different training schemes trained on Ising data, and the bottom figure shows these training metrics from machines trained on astrocytes across a range of training parameters, using CD, PT1, or PT5 training.

nearest-neighbour coupling only. I then trained ten RBMs using the CD algorithm, and a collection of RBMs using parallel tempering with K chains that linearly interpolate between $\beta = 0$ and $\beta = 1$, for $K = 1, 2, 3, 4, 10$, training ten machines at each K. Since I was only interested in the rate of convergence, the machines were trained for just 450 epochs, and with a learning rate of 0.001 so that the weights do not change too much between updates. To compensate for the small learning rate, a small batch size of 40 was chosen. Comparisons of the log-likelihood are shown in Figure 2.40.

For the Ising model at $T = 1.8$, PT improved the training. PT1, i.e. PCD, performed slightly better than CD, but was less stable. Extending this to $K = 2$ parallel tempering did not improve the log-likelihood by much, as $PT2$ only adds an infinite temperature chain, *i.e.* uniform noise, to PCD. Starting from PT3, however, the log-likelihood became more stable, and reached a much higher value than in the CD case. At $T = 2.4$, (not shown) the effect of adding PT chains was not very obvious, but PT1 already performed better than CD.

To see if parallel tempering can also improve the log-likelihood of the machines trained on gene expression data, I trained machines with learning rates from $\{0.1, 0.01, 0.001\}$, hidden layer size from $\{25, 50, 75, 100\}$, using PT1 and PT5, with a batch size of 5.

The bottom of Figure 2.40 shows that across these training parameters, none of the log-likelihoods stabilised (larger batch sizes also did not improve the stability). Increasing the accuracy of the gradients therefore did not seem to stabilise the log-likelihood, which means that it is most likely caused by an error in the estimation of the partition function, not of the gradients.

2.4 Discussion

> Opaque learning systems may get us to Babylon, but not to Athens.
>
> Judea Pearl [40]

In this chapter, I estimated maximum entropy interactions in various data sets by training restricted Boltzmann machines. In the simplest setting—a pairwise Ising model on a toroidal lattice—this was already done in [59], and I was able to reproduce their results: the machines learned the value and the structure of the pairwise interactions as the log-likelihood increased. I then added third-order interactions, and while the log-likelihood decayed during training and any single machine often failed to learn particular 3-point interactions, the mean across 20 trained machines accurately reproduced both the triplet structure and value of the interactions. Tracking the mean value of each of the interaction orders allowed some insight into how the RBMs learn: linear terms grew first, but they could not capture correlation structure and early on in training were partially replaced by pairwise interactions, which in turn got partially replaced by 3-point interactions to capture the higher dependencies in the training data.

Similarly, when training on traits from the UK Biobank, I recovered biologically plausible couplings, even with a decaying log-likelihood. To recover the two modes present in the data, the RBMs had to encode critical dynamics, which hindered moment estimation because samples were not *i.i.d.* . It would be interesting to see if this is a general phenomenon, so training RBMs on a gene expression data set that contains *e.g.* multiple cell types could be an interesting future direction. The encoded distribution showed hints of power-law behaviour, and in [170] the authors argue that power-law scaling is a general property of biological systems, and a strong indication of criticality. The authors of [233] have shown, however, that such power-law scaling does not mean that the underlying dynamics are critical, but that Zipf-like behaviour arises very naturally in systems with unobserved variables. Furthermore, a Zipfian power-law is associated to critical dynamics because it is is indicative of a diverging specific heat. In [251], the authors show that the specific heat associated to the visible layer of a trained RBM is not necessarily related to critical dynamics underlying the training data. My results seem to corroborate [233, 251], as the 67 traits most likely share many unobserved common causes, and there is no reason to assume that the 67 traits form a meaningful dynamical system. It has also been observed that very sparse binary data (where 1s are rare) can be accurately modelled by a critical Ising model [178].

I then trained 20 machines on gene expression data from embryonic mouse astrocytes. As before, I found that the log-likelihood was unstable, but machines accurately reproduced up to six moments of the data, and the encoded couplings had biological plausibility.

There were three main problems with the estimated interactions. First, the fact that the log-likelihood decayed made evaluating the trained RBMs and measuring out-of-sample generalisation or overfitting difficult. I tried to stabilise the log-likelihood by changing to a training scheme that used parallel tempering to estimate the gradients, but while this worked for pairwise Ising models, it did not stabilise the log-likelihood when training on gene expression data. This is an indication that the log-likelihood decay was not the result of inaccurate gradients, but rather a systematic error in the estimation of the partition function. The instability in the log-likelihood was present for many settings of the training parameters, but a more systematic grid search across all parameters, or a full calculation of the log-likelihood could give more insight into this in the future. However, while inconvenient, this estimation error does not directly affect the training procedure, and both these options are computationally expensive, so it was not considered crucial to the current research. As an alternative training metric, I estimated the moments of the sampling distribution from the RBMs, but this only showed that the RBMs were able to accurately reproduce all single-variable marginals. It would therefore be of interest to extend this analysis to cross-moments, *i.e.* the off-diagonal elements of the co-moments, to see if the joint distributions could also be reproduced.

The second problem with the interaction estimates from the RBMs is that they do not come with a level of confidence or uncertainty. There are various possibilities to assign significance to the interactions, like training multiple machines and fitting a distribution to each interaction estimate. However, there is no clear null hypothesis. I have tried to construct null hypotheses by training machines on different kinds of shuffled data, but none of these could be motivated from first principles. Furthermore, QQ plots of the resulting p-values led to almost all interactions being significant, a reflection both of a faulty null and the fact that the interactions are not independent estimates. Because none of these tests could be justified as yielding meaningful levels of confidence, I have decided not to include them in this thesis.

The third major problem with these estimates, most visible in the estimates of genetic interactions from gene expression data, is that they are not robust with respect to changing the number of variables in the training data. I found the mean across the 20 machines to be reasonably robust with respect to subsets of cells and subsets of genes, but upon including an additional 100 highly variable genes in the analysis, the estimates drastically changed. This revealed a strong omitted variable bias—as I included more genes, the machines restructured the network of dependencies, making it impossible to decide which interactions to assign biological meaning to.

These last two flaws are related, since perhaps a confidence level could determine which estimates are most robust. One way to generate statistically sound confidence intervals is by bootstrap resampling the data, and estimating the interactions in each bootstrap resample. However, this would mean training an RBM separately on each of the thousands of resamples, and is thus computationally intractable. The omitted-variable bias could be mitigated by training on many more genes, but this is also computationally expensive, and practically difficult without a good metric to evaluate training. It is for this reason that another approach is required, which is the subject of the rest of this thesis.

One area in which the trained RBMs could still serve a purpose, however, is in data generation. Once trained, the RBMs offer a computationally cheap way to generate samples from the encoded distribution. In particular, by fixing some of the visible nodes to be in a particular state, the RBM can generate samples from all conditional distributions in a process known as inpainting. It is not *a priori* obvious that samples generated in this way accurately reproduce the conditional distribution, but in Section 2.D I present a proof based on the do-calculus that under the Gibbs sampling procedure used in this chapter, the do-operator and the see-operator are the same, and fixing nodes indeed makes Gibbs samples reproduce the conditional distribution.

2.A Spin flips in the 3-point Ising model

Consider the system with only a 3pt coupling in the $\{0,1\}$ basis. To be able to use `Magneto` to simulate this system, this interaction should be transformed into the $\{-1,1\}$ basis. Note the following:

$$H = \sum_{<ijk>} J^{(3)} v_i v_j v_k \Big|_{\{0,1\}} \tag{2.32}$$

$$= \sum_{i,j,k} J_{ijk} \; v_i v_j v_k \Big|_{\{0,1\}} \tag{2.33}$$

Where $J_{ijk} = 0$ if (i,j,k) are not a straight connected triplet, and $J_{ijk} = J^{(3)}/6$ if they are.

$$= \sum_{i,j,k} J_{ijk} \left(\frac{v_i+1}{2}\right)\left(\frac{v_j+1}{2}\right)\left(\frac{v_k+1}{2}\right)\Big|_{\{-1,1\}} \tag{2.34}$$

$$= \sum_{i,j,k} \frac{J_{ijk}}{8} \left(v_i v_j v_k + v_i v_j + v_i v_k + v_j v_k + v_i + v_j + v_k\right)\Big|_{\{-1,1\}} + \text{constant} \tag{2.35}$$

Starting from Equation (2.35), the effect of a spin flip at site n is then written, without contracting anything, as

$$\Delta E_n = H_{v_n \to -v_n} - H \tag{2.36}$$

$$= \sum_{i,j,k \neq n} \frac{J_{ijk}}{8}\Big(v_i v_j v_k + v_i v_j + v_i v_k + v_j v_k + v_i + v_j + v_k\Big) \tag{2.37}$$

$$+ \sum_{j,k} \frac{J_{njk}}{8}\Big(- v_n v_j v_k - v_n v_j - v_n v_k + v_j v_k - v_n + v_j + v_k\Big) \tag{2.38}$$

$$+ \sum_{i,k} \frac{J_{ink}}{8}\Big(- v_i v_n v_k - v_i v_n + v_i v_k - v_n v_k + v_i - v_n + v_k\Big) \tag{2.39}$$

$$+ \sum_{i,j} \frac{J_{ijn}}{8}\Big(- v_i v_j v_n + v_i v_j - v_i v_n - v_j v_n + v_i + v_j - v_n\Big) \tag{2.40}$$

$$- \sum_{i,j,k} \frac{J_{ijk}}{8}\Big(v_i v_j v_k + v_i v_j + v_i v_k + v_j v_k + v_i + v_j + v_k\Big) \tag{2.41}$$

The full first sum $\sum_{i,j,k\neq n}$ cancels against the part of the last sum where $i, j, k \neq n$. On top of that, all terms in the spin-flipped sums that did not contain a v_n did not switch sign so they also cancel against the last sum. From this last sum, only the terms where i, j or k is n and that contain a v_n are left, which can just be added to the remaining terms:

$$= -2\sum_{j,k}\frac{J_{njk}}{8}\Big(v_n v_j v_k + v_n v_j + v_n v_k + v_n\Big) \tag{2.42}$$

$$- 2\sum_{i,k}\frac{J_{ink}}{8}\Big(v_i v_n v_k + v_i v_n + v_n v_k + v_n\Big) \tag{2.43}$$

$$- 2\sum_{i,j}\frac{J_{ijn}}{8}\Big(v_i v_j v_n + v_i v_n + v_j v_n + v_n\Big) \tag{2.44}$$

As J_{ijk} is symmetric, these terms can be brought under one summation:

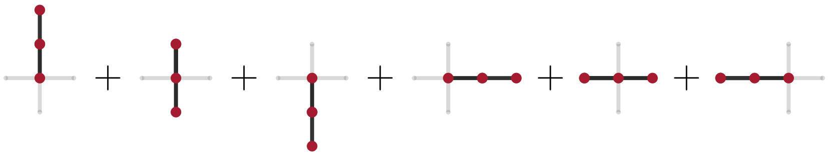
$$= -2\sum_{j,k}\frac{3\cdot J_{njk}}{8}\Big(v_n v_j v_k + v_n v_j + v_n v_k + v_n\Big) \tag{2.45}$$

Or rather, having the sum go over all (j, k) that form a straight connected triplet with n:

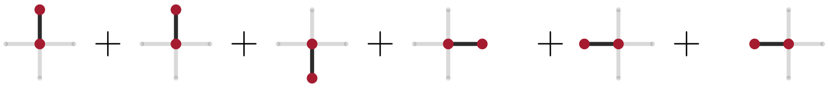
$$= -2\sum_{(j,k)\in<njk>}\frac{3\cdot J^{(3)}}{8}\Big(v_n v_j v_k + v_n v_j + v_n v_k + v_n\Big) \tag{2.46}$$

The (j, k) that are included in this sum can be written out diagrammatically. In the diagrams, site n is always in the centre, and the order of the terms $(j, k) \in< njk >$ is as in the previous diagrammatic expansion of ΔE.

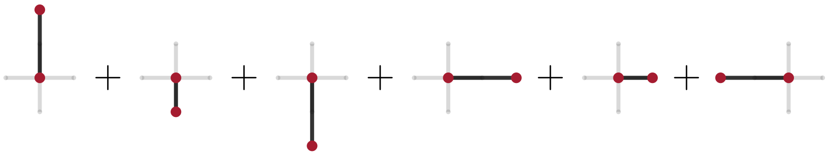
$$\sum_{(j,k)\in<njk>} v_n v_j v_k =$$



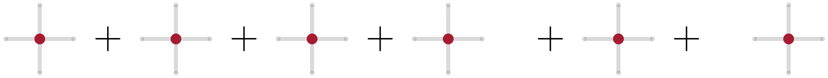
$$\sum_{(j,k)\in<njk>} v_n v_j =$$



$$\sum_{(j,k)\in<njk>} v_n v_k =$$



$$\sum_{(j,k)\in<njk>} v_n =$$



This leads to the conclusion that a spin flip in a magneto run should affect:

- Six 3-point couplings with weight $3 \cdot J^{(3)}/8$

- Four nearest neighbour couplings with weight $2 \cdot 3 \cdot J^{(3)}/8$
- Four next-to nearest neighbour couplings with weight $3 \cdot J^{(3)}/8$
- One linear term with weight $6 \cdot 3 \cdot J^{(3)}/8$

Which is how I have implemented the metropolis sampling.

2.B The machines learn 3-point interactions added to the astrocyte data

The Ising experiments showed that the RBMs can learn the structure of 3-point couplings. I verified that this ability persists in the context of real biological data by adding 5 artificial variables to the 37 genes. Of these five variables, three coupled with only a pure triplet interaction as in Section 2.3.1, and two did not couple at all. Figure 2.41 shows that the RBMs correctly identified the 3-point interaction. As in Section 2.3.1, the RBMs included a small spurious 2-point coupling between two of the genes that in reality couple with a three-point coupling, but the signal was much weaker than the signal for the 3-point coupling. Figure 2.42 shows the extracted couplings for each machine, a comparison of the three-point couplings that should be found with the ones that should not, and the pairwise interactions between the relevant genes.

2.C The KL-divergence does not add information beyond the log-likelihood

Before training machines myself, I analysed the machines that are presented in the paper [59], trained on 8 × 8 lattices, as they were on Eddie on May 8th 2019. The authors already illustrated that the training was stable and accurate, but I quantified this accuracy in terms of the KL-divergence between the empirical data distribution p_{data} and the distribution that the trained RBMs encoded in their visible layer p_{RBM}. The KL-divergence can be expressed as follows:

$$D_{\text{KL}}(p_{\text{data}} \,||\, p_{\text{RBM}}) = \sum_{v} \; p_{\text{data}}(v) \log\Big(p_{\text{data}}(v)\Big) - p_{\text{data}}(v) \log\Big(p_{\text{RBM}}(v)\Big) \tag{2.47}$$

$$= -S_{\text{data}} - LL_{\text{RBM}} \tag{2.48}$$

This shows that on a data set of a given entropy, the log-likelihood LL_{RBM} of observing the data in the RBMs visible layer is indeed a direct measure of how accurately the data distribution is reproduced, but that to compare accuracy across entropies, *i.e.* temperatures, one should subtract the data entropy. The sum over v sums over all possible states of the visible layer, which is intractable for an 8 × 8 lattice, has to be approximated. States for which p_{data} is large will contribute most to the sum and appear most in the training data. I assumed that these states will appear sufficiently in the training data for their probability to be reasonably well approximated by $n(v)/N$, where $n(v)$ is the number of occurrences of state v, and $N = \sum_v n(v)$ is the total number of states observed. Using this approximation, and the log-likelihood at the end of training,

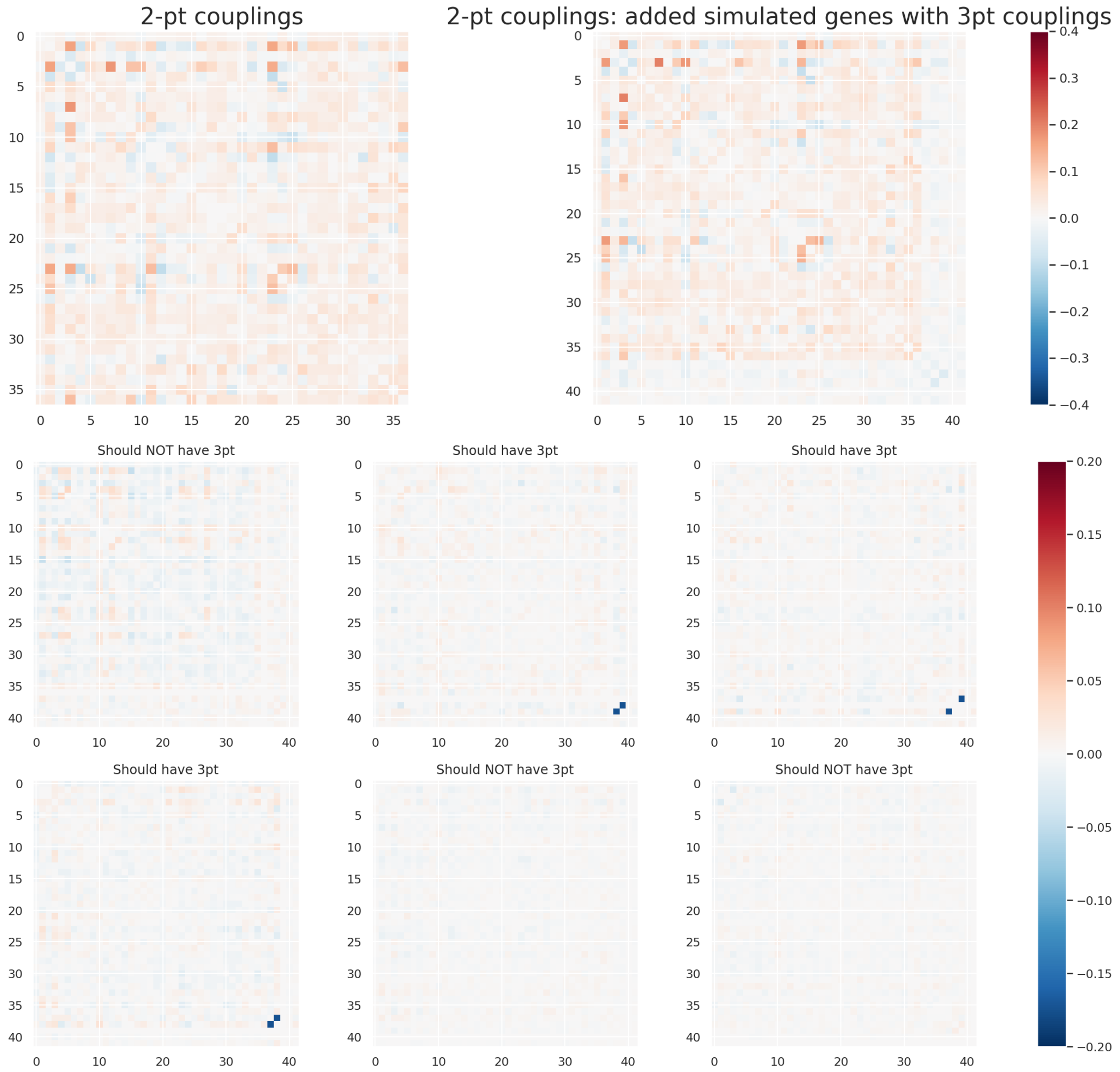


Figure 2.41: **Top:** The 2-point coupling matrix of the original data, and the data with simulated three-point couplings. **Bottom:** Slices of the three-point coupling tensor show that an artificially added 3-point interaction is accurately reproduced.

the KL-divergence can be calculated for any given machine. Reported in Figure 2.43 is the KL-divergence for the machines from the Ising paper. The shown log-likelihood is the mean of the 100 last machines (1k epochs). Vertical error bars that show the standard deviation of the log-likelihood across these 100 machines are shown, but are too small to be visible.

In Figure 6 of [59], it can be seen that as the temperature increased, the estimated first moment, the magnetisation, became less accurate, while the second moment, the susceptibility, became more accurate. The temperature-KL-divergence curve in Figure 2.43 shows that overall, the machines trained best at a temperature of 1.9. This result does come with the caveat that the data entropy was approximated, which could distort the results. In particular, if at any point there are no states that appear twice, then that distribution will have the theoretically maximal entropy, which is indeed achieved for temperatures above 2.6. Because of this phenomenon,and the fact that the KL-divergence does not add much insight into the training accuracy beyond what the log-

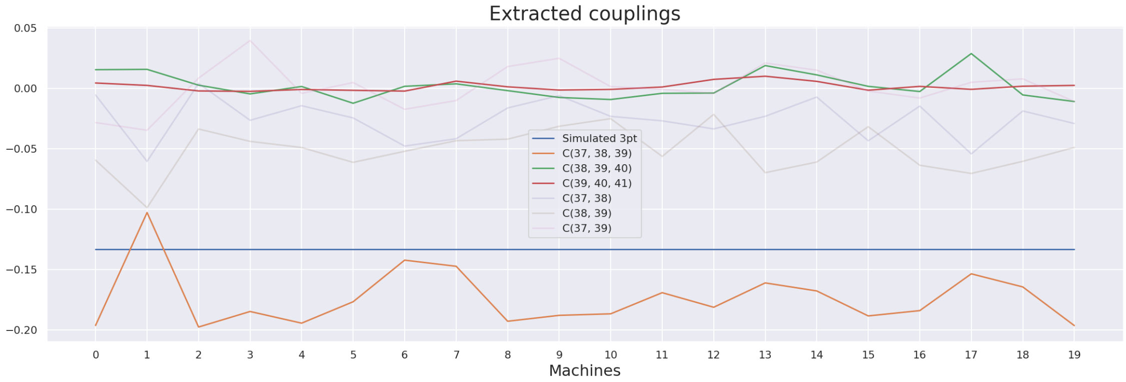
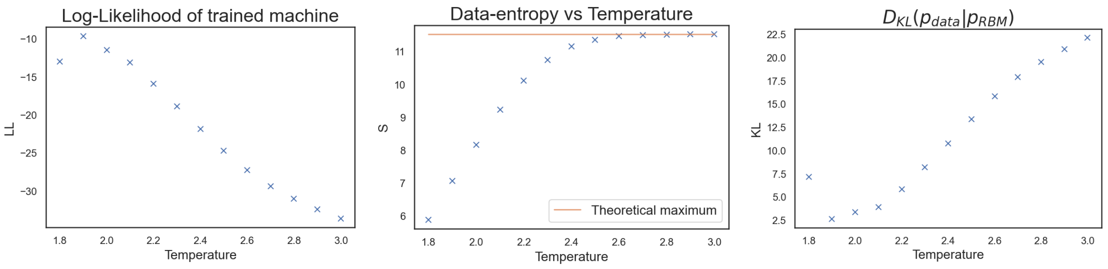


Figure 2.42: The 3-point interactions that should be zero, were zero. The 3-point that should be $J^{(3)}/(3T) = -0.133$ is -0.174 ± 0.02, so the triplet slightly overcouples. The 2-point interactions were all smaller than the 3-point interaction amongst the interacting triplet.



(a)

(b)

(c)

Figure 2.43: Properties of the training data and trained machines for different temperatures.

likelihood provided, I focused on the log-likelihood throughout this chapter.

2.D The see- and do-operator are the same in RBMs

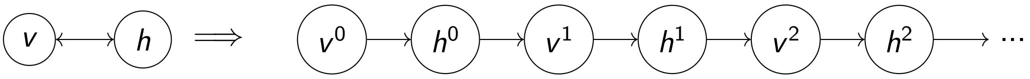
Given a Bayesian network represented by a DAG $G = (V, E)$, the fundamental problem in causal inference is to estimate quantities like $p_G\,(V_A = v_a \mid do(V_B = v_b))$ for an arbitrary partition of the vertices $V = V_A \cup V_B$. The do-calculus provides rules and methods for answering such questions [190]. Intervening on an RBM by fixing some of the visible nodes to certain values is often done to sample from the conditional distribution—a process known as inpainting [81]—but I found no proof that this do-operation is equivalent to sampling from the conditional distribution. In fact, the standard formulation of an RBM is not suitable for treatment by the do-calculus, as the network of dependencies does not form a DAG. However, by exploiting the sequential nature of Gibbs-sampling, I show the following:

Lemma 1 (Inpainting is conditional sampling)**.** *Consider a restricted Boltzmann machine with visible nodes v and hidden nodes h, where $|v| > 0$ and $|h| > 0$. Let $p(v)$ be the marginal probability distribution over the visible nodes, and $v = v_a \cup v_b$ an arbitrary* *partition of the visible nodes. Consider the nodes of the RBM as random variables that evolve under alternating Gibbs sampling of the visible and the hidden layers. Define the do-operator on a variable X as fixing that variable $X = x$ before generating each Gibbs sample. Then,*

$$p(v_a \mid do(v_b)) = p(v_a \mid see(v_b)) = p(v_a \mid v_b) \tag{2.49}$$

That is, inpainting is sampling from the conditional distribution.

Proof. First note that the RBM has to be represented by a DAG. To do this, consider the nodes as random variables that evolve under Gibbs sampling. The bipartite structure of the network allows us to unroll the network in time:



where v^i and h^i mark the i'th Gibbs sample of the visible and hidden layer, respectively. Given the partition $v = v_a \cup v_b$, the following should be verified:

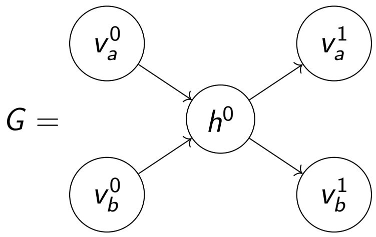
$$p(v_a^1 \mid do(v_b^0)) = p(v_a^1 \mid v_b^0) \tag{2.50}$$

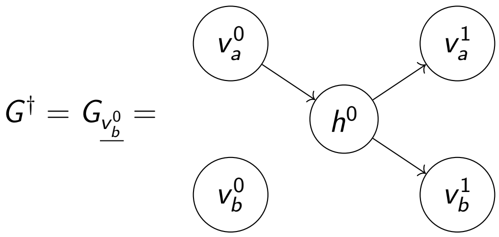
The second rule of the do-calculus states [190]:

Rule 2 (Action/observation exchange)

$$p(y \mid do(x), do(z), w) = p(y \mid do(x), z, w) \text{ if } (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}\underline{Z}}} \tag{2.51}$$

Where $G_{\overline{X}\underline{Z}}$ is the graph G with all the arrows into X, and out of Z removed. If $X = W = \emptyset$, then Rule 2 can be directly apply to Equation (2.50) by defining the following two graphs:





where v could be partitioned because at any given time point the visible nodes are mutually independent conditional on the hidden layer. The value of v_b^1 will be discarded and set to v_b^0 again, but that is irrelevant in the present discussion. Now $(v_a^1 \perp\!\!\!\perp v_b^0)_{G^\dagger}$, which by **Rule 2** implies Equation (2.50), and completes the proof.

□

Example As a very simple example, consider the case where $v_a = v_1$, $v_b = v_2$ and h_1 all comprise just a single node. The full distribution is:

$$P_G(v_1, v_2, h_1) = \frac{1}{\mathcal{Z}_G} e^{h_1 w_{11} v_1 + h_1 w_{12} v_2 + b_1 v_1 + b_2 v_2 + c_1 h_i} \tag{2.52}$$

Denoting by (abc) the situation in which $v_1 = a,\ v_2 = b,\ h_1 = c$, the conditional distribution after observing $v_2 = 1$ is:

$$
\begin{aligned}
P(v_1 = 1|\text{see}(v_2 = 1)) &= \frac{P_G(v_1 = 1, v_2 = 1)}{P_G(v_2 = 1)} && (2.53)\\
&= \frac{(110) + (111)}{(111) + (110) + (011) + (010)} && (2.54)\\
&= \frac{e^{b_1+b_2} + e^{w_{11}+w_{12}+b_1+b_2+c_1}}{e^{w_{11}+w12+b_1+b_2+c_1} + e^{b_1+b_2} + e^{w_{12}+b_2+c_1} + e^{b_2}} && (2.55)\\
&= \frac{e^{b_1} + e^{w_{11}+w_{12}+b_1+c_1}}{e^{w_{11}+w12+b_1+c_1} + e^{b_1} + e^{w_{12}+c_1} + 1} && (2.56)
\end{aligned}
$$

Now consider the intervention $do(v_2 = 1)$. It just adds a bias w_{12} to the hidden layer. Denoting by (ab) that $v_1 = a,\ h_1 = b$:

$$
\begin{aligned}
P_G(v_1 = 1|\text{do}(v_2 = 1)) &= P_{G^\dagger}(v_1 = 1) && (2.57)\\
&= \frac{1}{\mathcal{Z}_{G^\dagger}}\Big((11) + (10)\Big) && (2.58)
\end{aligned}
$$

Writing out this partition function:

$$
\begin{aligned}
Z_{G^\dagger} &= (11) + (10) + (01) + (00) && (2.59)\\
&= e^{w_{11}+b_1+c_1+w_{12}} + e^{b_1} + e^{c_1+w_{12}} + 1 && (2.60)
\end{aligned}
$$

So that

$$
P(v_1 = 1|\text{do}(v_2 = 1)) = \frac{e^{w_{11}+b_1+c_1+w_{12}} + e^{b_1}}{e^{w_{11}+b_1+c_1+w_{12}} + e^{b_1} + e^{c_1+w_{12}} + 1} \qquad (2.61)
$$

which indeed coincides with Equation (2.56).

2.E Full list of UK Biobank traits

0	f05 delirium	34	arm predicted mass (left)
1	n20-n23 urolithiasis	35	reticulocyte percentage
2	n20 calculus of kidney and ureter	36	high light scatter reticulocyte count
3	h60 otitis externa	37	arm fat-free mass (left)
4	g35 multiple sclerosis	38	hip circumference
5	d05 carcinoma in situ of breast	39	trunk predicted mass
6	j45 asthma	40	trunk fat-free mass
7	inflammatory bowel disease	41	arm predicted mass (right)
8	irritable bowel syndrome	42	basal metabolic rate
9	m86-m90 other osteopathies	43	mean corpusc. haemoglobin concentration
10	m05 seropositive rheumatoid arthritis	44	arm fat-free mass (right)
11	rheumatoid arthritis	45	trunk fat percentage
12	k50 crohns disease [regional enteritis]	46	eosinophill count
13	d04 carcinoma in situ of skin	47	body fat percentage
14	depression	48	high light scatter reticulocyte %
15	d80-d89 disorders involving the immune mech.	49	monocyte percentage
16	d86 sarcoidosis	50	hand grip strength (right)
17	k58 irritable bowel syndrome	51	whole body fat-free mass
18	e10-e14 diabetes mellitus	52	haematocrit percentage
19	c61 malignant neoplasm of prostate	53	whole body water mass
20	high cholesterol	54	leg fat mass (right)
21	white blood cell (leukocyte) count	55	leg predicted mass (left)
22	neutrophill count	56	leg fat-free mass (left)
23	standing height	57	leg fat percentage (right)
24	platelet crit	58	hand grip strength (left)
25	lymphocyte count	59	number of vehicles in household
26	monocyte count	60	skin colour
27	platelet count	61	tea intake
28	sitting height	62	Female
29	trunk fat mass	63	Male
30	reticulocyte count	64	Age: 40-50
31	weight	65	Age: 50-60
32	whole body fat mass	66	Age: 60-70
33	weight		

Chapter 3

Model-free interactions: Theory and simulations

> The simulacrum is never what hides the truth—it is truth that hides the fact that there is none. The simulacrum is true.
>
> Baudrillard [23]

> Hypotheses non fingo.
>
> Isaac Newton [177]

䷫

3.1 Introduction

Learning from data involves three mathematical objects that are often conflated, but fundamentally different: the quantity one wants to learn about, the estimator used to do so, and the generated estimate. For example, a quantity of interest could be the interactions among a set of variables $\mathbf{X}$ and an outcome $\mathbf{y}$. This can be modelled with a linear model $\mathbf{X}\beta = \mathbf{y}$ where the quantity of interest is identified with β. Estimating β is usually done with a least-squared *estimator* as $\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, where the hat reflects the fact that $\widehat{\beta}$ is an estimator for the quantity β. The *estimate* then depends on what realisation $\mathbf{y}$ is used to calculate $\widehat{\beta}$.

In Chapter 2 of this thesis, the quantity of interest was the set of maximum-entropy interactions among a collection of variables. The estimator I used was the training procedure of a restricted Boltzmann machine, and the estimate was encoded in the final state of the machine. It turned out that this estimator was not very stable: there were no perfect training metrics, estimates differed upon retraining, were not robust to variable selection, and including more variables made training intractable. In the second part of my research, I used a different estimator to estimate the interactions. The estimator I used was introduced in [24], and I will refer to it as a *model-free interaction* (MFI). To justify its use, I will first highlight the problematic role of a model in an estimation procedure.

3.1.1 Model bias

> When a measure becomes a target, it ceases to be a good measure.
>
> Goodhart's law

Often, the quantity of interest is a mathematically ambiguous object. For example, *'the interaction between gene A and gene B'* does not specify what is actually meant by *'interaction'*. To make this explicit, one could write down a model that describes the system, and specify which parameter corresponds to the interaction. Consider the linear model $\mathbf{X}\beta = y$ that describes how the scalar outcome y depends on the quantities $\mathbf{X}$ through the parameters β. The estimate of $\widehat{\beta}$ depends on which quantities are included in $\mathbf{X}$, the design matrix. Let the ground truth be the simplest case of a bilinear interaction with Gaussian additive noise: $y = x_1 + x_2 + x_1x_2 + \eta$, where $\eta \sim \mathcal{N}(0, 0.1)$. I generated 1,000 samples from this model, sampling x_1 and x_2 from the uniform distribution over the interval $[-1, 1]$. I then fitted five different models for y to this data with increasingly high powers of x_1 in the interaction term: $y = \mathbf{X}\beta = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^n x_2$ for $n = 1, \dots, 5$, yielding five different estimates for the interaction parameter β_3, shown in Figure 3.1. Only the model that exactly corresponds to the ground truth yielded an interaction estimate within statistics of the ground truth. Since the sampling distribution of x_1 and x_2 was an even function, models that introduced an interaction with an even power of x_1 have interaction estimates centred around 0—they incorrectly conclude there is no interaction. Models with odd powers of x_1 in the interaction term increasingly overestimate the interaction. This is not a reflection of an inherent bias of the OLS estimator for parameters of a linear model (OLS is an unbiased estimator for these parameters). Rather, it is a reflection of the fact that the model is misspecified, which results in a *model bias* in the interaction estimates. The estimate is still unbiased, but estimates something different from the quantity of interest.

There is another kind of model bias that results from a particular kind of misspecification, namely omitting a variable in the model. This happens, for example, when it is incorrectly assumed that a particular variable is irrelevant to the outcome, or if there is no data available on the value of that particular variable. As an example, consider the case in which the ground truth is described by $y = x_1 + x_2 + x_3 + x_1x_2 - x_1x_2x_3 + \eta$. I generated 1,000 samples from this model, sampling x_1, x_2 and x_3 from the uniform distribution over the interval $[0, 3]$. I then fitted two models to the data:

$$y^{(1)} = \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_{12}x_1x_2 + \beta_{123}x_1x_2x_3$$
$$y^{(2)} = \beta_1x_1 + \beta_2x_2 + \beta_{12}x_1x_2$$

The estimates for β_{12} are shown in Figure 3.2. Omitting x_3 resulted in an interaction estimate that disagreed with the ground truth not just in value, but even in sign.

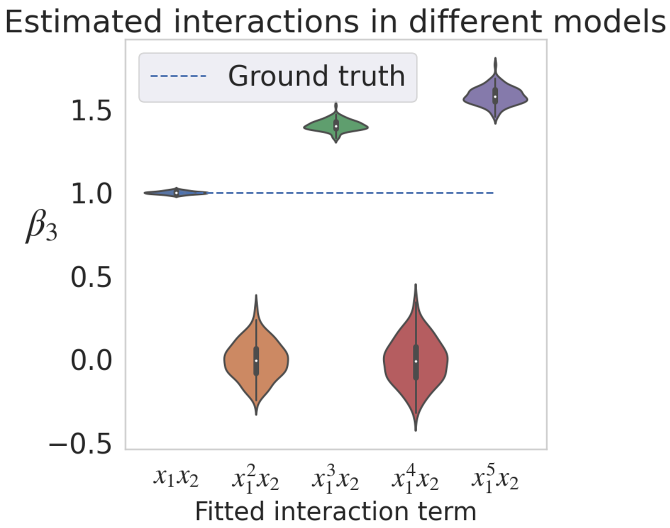


Figure 3.1: Five different estimates of the interaction between x_1 and x_2. Only the model that exactly corresponds to the ground truth (the bilinear interaction term x_1x_2) yielded an interaction estimate within statistics of the ground truth. The estimates were based on 1,000 samples from y that were generated by sampling x_1 and x_2 from the uniform distribution on $[-1, 1]$. The estimation was repeated 100 times to get a range of values for each estimate.

It is rare to know the ground-truth model, so almost any model-based estimate will suffer from model bias. It is therefore desirable to be able to specify the quantity of interest directly in terms of the data in a precise way, bypassing the need to introduce a model. An example of such a *model-free* quantity is Pearson correlation. It is defined on two random variables X and Y as $\rho_{XY} = \frac{cov(X,Y)}{\sigma_X\sigma_Y}$, where σ_Z is the standard deviation of a variable Z. There is an unbiased estimator on a sample of N observations (x_i, y_i) for

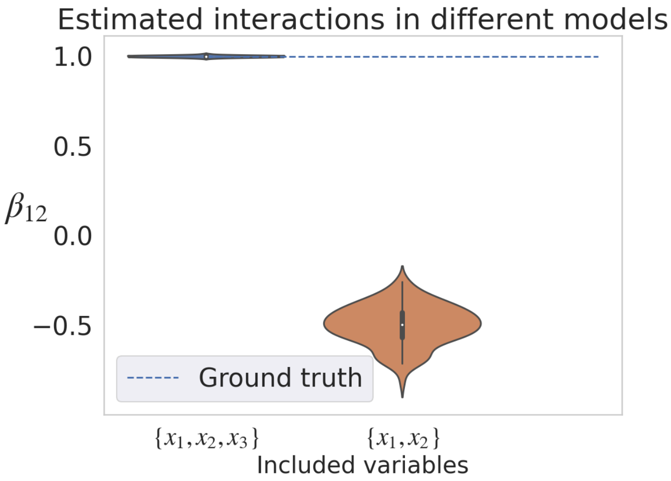


Figure 3.2: Two different estimates of the interaction between x_1 and x_2. Omitting x_3 from the model makes the interaction estimate incorrect in both value and sign. The estimates were based on 1,000 samples from y that were generated by sampling x_i from the uniform distribution on $[0, 3]$. The estimation was repeated 100 times to get a range of values for each estimate.

this:

$$\widehat{\rho}_{XY} = \frac{\sum_{n=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{n=1}^{N}(x_i - \overline{x})^2}\sqrt{\sum_{n=1}^{N}(y_i - \overline{y})^2}} \tag{3.1}$$

where $\overline{z}$ denotes the sample mean of a variable z. This does not require assumptions about the functional form of the relationship between X and Y, and is thus a model-free estimator. A model-free estimate just is what it is: there is no ground truth to compare it with. This is a desirable property of an estimator because it removes the need to justify a model, but it comes at the cost of being harder to interpret as it does not correspond to a model parameter with a certain fixed interpretation. In this chapter, I will study a model-free definition of interactions that coincides with the Ising interactions explored in Chapter 2.

3.1.2 Aim and outline of this chapter

This chapter aims to develop a deeper understanding and intuition for model-free interactions, which will then be applied to gene expression data in **Chapters 4 and 5**. **Section 3.2.1** introduces a definition of MFIs based on how they were introduced in [24]. **Section 3.2.2** contains more details and some new results concerning the practical estimation procedure. One of the most important advantages of MFIs over the RBM estimates is that they can be assigned a confidence level as outlined in **Section 3.2.3**. **Chapter 3.3** contains a purely theoretical result that relates the MFIs to information theory (**Section 3.3.1**), and the first calculations of MFIs on logic gates (**Section 3.3.2**) and on simulated data from various causal dynamics (**Section 3.3.3**).

Parts of this chapter have previously appeared in a preprint [125] and subsequent publication [126].

3.2 Methods

3.2.1 Model-free definition of interaction

> The elementary unit of information is a difference which makes a difference.
>
> Gregory Bateson [20]

Let us take a step back: What do we mean when we say *interaction*? An interaction among n variables describes how the joint state of the n variables influences a particular outcome *beyond the influence of the marginal states*. In other words: an interaction corresponds to a change in the effect of each of the variables on the outcome, determined by the joint state of the other variables. This definition of interaction was made precise in [24] (there, these are called the *additive* interactions), and an equivalent definition follows here.

The isolated effect, or *1-point interaction*, $I_i^{(Y)}$ of a variable $X_i \in X$ on an observable Y is defined as the partial derivative of Y along X_i:

$$I_i^{(Y)} = \left.\frac{\partial Y}{\partial X_i}\right|_{\underline{X}=0} \quad , \quad \underline{X} = X \setminus \{X_i\} \tag{3.2}$$

where the effect of X_i on Y is isolated by conditioning on all other variables being zero. This expression is well-defined as the restriction of a derivative is the derivative of the restriction. A pair of variables X_i and X_j has a 2-point interaction $I_{ij}^{(Y)}$ when the value of X_j changes the isolated effect of X_i on Y:

$$I_{ij}^{(Y)} = \left.\frac{\partial I_i^{(Y)}}{\partial X_j}\right|_{\underline{X}=0} = \left.\frac{\partial^2 Y}{\partial X_j \partial X_i}\right|_{\underline{X}=0} \quad , \quad \underline{X} = X \setminus \{X_i, X_j\} \tag{3.3}$$

A third variable X_k can modulate this interaction through what is called a 3-point interaction $I_{ijk}^{(Y)}$:

$$I_{ijk}^{(Y)} = \left.\frac{\partial I_{ij}^{(Y)}}{\partial X_k}\right|_{\underline{X}=0} = \left.\frac{\partial^3 Y}{\partial X_k \partial X_j \partial X_i}\right|_{\underline{X}=0} \quad , \quad \underline{X} = X \setminus \{X_i, X_j, X_k\} \tag{3.4}$$

This process of taking derivatives with respect to an increasing number of variables can be repeated to define n-point interactions:

Definition 4 (n-point interaction with respect to outcome Y)**.** *Let p be a probability distribution over a set of random variables $X = \{X_i \mid 1 \leq i \leq N\}$, taking values in $\mathcal{X}$. Let Y be a function $Y : \mathcal{X} \to \mathbb{R}$, differentiable over $\mathcal{X}$, and let n be a natural number $1 \leq n \leq |X|$. Then the n-point interaction among the variables $\{X_1, \dots, X_n\} \subseteq X$ with respect to the outcome Y is written as $I_{X_1 \dots X_n}^{(Y)}$ and given by*

$$I_{X_1 \dots X_n}^{(Y)} = \left.\frac{\partial^n Y(X)}{\partial X_1 \dots \partial X_n}\right|_{\underline{X}=0} \tag{3.5}$$

where $\underline{X} = X \setminus \{X_1, \dots X_n\}$.

This definition of interaction makes explicit the fact that interactions are defined with respect to a particular outcome. I follow [24] and define interactions with respect to the most general outcome: the (log of the) joint distribution $p(X)$ over all variables X.

Definition 5 (Model-free n-point interaction). *A model-free n-point interaction (MFI) is an n-point interaction among binary random variables with respect to the logarithm of their joint probability:*

$$I_{1\dots n} := I^{(\log p(X))}_{X_1\dots X_n} = \frac{\partial^n \log p(X)}{\partial X_1 \dots \partial X_n}\bigg|_{\underline{X}=0} \tag{3.6}$$

where $\underline{X} = X \setminus \{X_1, \dots X_n\}$.

In [24], the authors noted that when the variables X_i are restricted to binary values—*i.e.* $\mathcal{X} = \{0,1\}^{|X|}$—n-point interactions become model-free in the sense that they are ratios of probabilities that do not involve the functional form of the joint probability distribution. To see this, note that for any function $f : \mathbb{B} \to \mathbb{R}$, where $\mathbb{B}$ is the set of Boolean values $\{0,1\}$, the Boolean derivative is just a difference:

$$\frac{\partial}{\partial x} f(x) = f(1) - f(0) \tag{3.7}$$

such that the model-free 1-, 2-, and 3-point interactions can be written as:

$$\begin{aligned}
I_i &= \frac{\partial}{\partial X_i} \log p(X_i \mid \underline{X}) && (3.8)\\
&= \frac{p\left(X_i = 1 \mid \underline{X} = 0\right)}{p\left(X_i = 0 \mid \underline{X} = 0\right)} && (3.9)\\
I_{ij} &= \frac{\partial^2}{\partial X_i \partial X_j} \log p(X_i, X_j \mid \underline{X}) && (3.10)\\
&= \frac{p\left(X_{ij} = (1,1) \mid \underline{X} = 0\right) p\left(X_{ij} = (0,0) \mid \underline{X} = 0\right)}{p\left(X_{ij} = (1,0) \mid \underline{X} = 0\right) p\left(X_{ij} = (0,1) \mid \underline{X} = 0\right)} && (3.11)\\
I_{ijk} &= \frac{\partial^3}{\partial X_i \partial X_j \partial X_k} \log p(X_i, X_j, X_k \mid \underline{X}) && (3.12)\\
&= \frac{p\left(X_{ijk} = (1,1,1) \mid \underline{X} = 0\right) p\left(X_{ijk} = (1,0,0) \mid \underline{X} = 0\right)}{p\left(X_{ijk} = (1,1,0) \mid \underline{X} = 0\right) p\left(X_{ijk} = (1,0,1) \mid \underline{X} = 0\right)} && (3.13)\\
&\times \frac{p\left(X_{ijk} = (0,1,0) \mid \underline{X} = 0\right) p\left(X_{ijk} = (0,0,1) \mid \underline{X} = 0\right)}{p\left(X_{ijk} = (0,1,1) \mid \underline{X} = 0\right) p\left(X_{ijk} = (0,0,0) \mid \underline{X} = 0\right)} && (3.14)
\end{aligned}$$

where $X_{i\dots k}$ denotes the tuple $(X_i, \dots, X_k)$ and Bayes' rule was used to replace joint with conditional probabilities. This definition of interaction has the following properties:

- It is symmetric in the variables: $I_S = I_{\pi(S)}$ for any set of variables S, and any permutation π.
- Conditionally independent variables do not interact: $X_i \perp\!\!\!\perp X_j \mid \underline{X} \implies I_{ij} = 0$.
- If $\underline{X} = \emptyset$, the definition coincides with that of a generalised log-odds ratio, which has already been considered as an abstract notion of interaction in e.g. [96] and [19].

- The interactions are model-free: no knowledge of the functional form of $p(X)$ is required, and the probabilities can be directly estimated from *i.i.d.* samples.
- If the ground truth is a Boltzmann distribution, then these interactions coincide exactly with the Ising, or maximum entropy, interactions from the previous Chapter 2.

A more thorough investigation of this definition and its relationship to information theory can be found in Section 3.3.1. For now, I will focus on the more practical issue of its estimation.

3.2.2 Estimation

Estimating the model-free interactions from Definition 5 on data involves estimating probabilities $p(X)$ of certain joint states X occurring. The true probabilities are usually unknown, but the interactions can be rewritten in terms of expectation values as follows. Note that all interactions involve factors of the type

$$\frac{p(X=1, Y=y \mid Z=0)}{p(X=0, Y=y \mid Z=0)} = \frac{p(X=1 \mid Y=y, Z=0)}{p(X=0 \mid Y=y, Z=0)} \tag{3.15}$$

$$= \frac{p(X=1 \mid Y=y, Z=0)}{1-p(X=1 \mid Y=y, Z=0)} \tag{3.16}$$

which can be written as

$$= \frac{\mathbb{E}[X \mid Y=y, Z=0]}{1-\mathbb{E}[X \mid Y=y, Z=0]} \tag{3.17}$$

since

$$\mathbb{E}[X \mid Z=z] = \sum_{x\in\{0,1\}} p(X=x \mid Z=z)\, x = p(X=1 \mid Z=z) \tag{3.18}$$

The 2-point interaction, for instance, can thus be written as

$$I_{ij} = \log \frac{\mathbb{E}\left(X_i | X_j=1, \underline{X}=0\right)\left(1-\mathbb{E}\left(X_i | X_j=0, \underline{X}=0\right)\right)}{\mathbb{E}\left(X_i | X_j=0, \underline{X}=0\right)\left(1-\mathbb{E}\left(X_i | X_j=1, \underline{X}=0\right)\right)} \tag{3.19}$$

An expectation value is still a parameter of the theoretical population, not an empirical sample statistic, but each expectation value in Equation (3.19) can be estimated from a sample in an unbiased way with sample means. However, the stringent conditioning in this estimator can make the number of samples that satisfy the conditioning very small, which gives the estimates a large variance (revealed upon bootstrap resampling, see Section 3.2.3). Note that if there is a subset of variables $\text{MB}_{X_i} \subseteq X$ such that $\forall X_k \in X \setminus (\text{MB}_{X_i} \cup X_i) : \; X_i \perp\!\!\!\perp X_k \mid \text{MB}_{X_i}$—in causal language: a set of variables MB_{X_i} that d-separates X_i from the rest—then one only has to condition on MB_{X_i} in Equation (3.19), reducing the variance of the estimator. Such a set MB_{X_i} is called a *Markov blanket*[1] of the node X_i. Since conditioning on fewer variables should reduce

[1] There has recently been some confusion around the notion of Markov blankets in biology, specifically with respect to their use in the so-called free energy principle. Throughout this thesis, *Markov blanket* refers to the notion of a *Pearl blanket* in the language of [41].

the variance of the estimator by increasing the number of samples that can be used for the estimation, one is generally interested in finding the smallest Markov blanket. The smallest Markov blanket is called the Markov boundary. However, in the absence of perfect knowledge of all conditional dependencies, the true Markov boundary is unknown, so throughout this thesis I will write Markov blanket to refer to the smallest blanket I could find. To further shrink the Markov blanket, note that the order of the indices in Equation (3.19) and its nth-order generalisation are arbitrary, so an n-point interaction has n equivalent forms—one using the expectation value of each of the n variables. Since smaller Markov blankets generally lead to smaller variance, estimating the interaction using the expectation value of the variable with the smallest Markov blanket will usually lead to the most precise estimation. However, a smaller Markov blanket is not *guaranteed* to lead to smaller variance, as the variance also depends on which states are actually present in the empirical distribution. Therefore, throughout this thesis, I will estimate each n-point interaction in n different ways, only reporting the most precise estimate—unless otherwise stated.

Finding such Markov blankets is hard. In fact, since it requires testing each possible conditional dependency between the variables, I claim (without proof) it is *Rung 2*-hard, referring to Pearl's ladder of causality [190]. That is, finding the smallest Markov blanket is at least as computationally complex as constructing a DAG of conditional dependencies consistent with the joint probability distribution, if such a graph exists. Finding the true Markov boundary among all variables is *Rung 3*-hard, as it requires knowledge of all confounders and causal relationships, but this is not necessary to estimate Equation (3.19).

Markov blankets are more than a computational trick—in theory, only variables that are in each other's Markov blanket can share a nonzero interaction. To see this, first note that the property of being in a variable's Markov blanket is symmetric:

Lemma 2 (Symmetry of Markov blankets)**.** *Let X be a set of variables with joint distribution $p(X)$. Let $A \in X$ and $B \in X$ such that $A \neq B$. Denote the minimal Markov blanket of X by MB_X. Then $A \in MB_B \iff B \in MB_A$, and we say that A and B are Markov-connected.*

I could not find a proof or reference for this basic fact, so I have included a proof in section 3.A. This definition of Markov-connectedness allowed me to state and prove the following (proof included in section 3.A):

Theorem 1 (Only Markov-connected variables interact)**.** *A model-free n-point interaction $I_{1\ldots n}$ can be nonzero if and only if all variables $S = \{X_1, \ldots, X_n\}$ are mutually Markov-connected.*

Knowledge of the causal graph thus helps estimation in two ways: it shrinks the variance of the estimator by relaxing the conditioning, and in addition identifies the interactions that could be nonzero.

Under imperfect knowledge of the causal graph, a variable might be accidentally excluded from the Markov blanket, which results in underconditioned probabilities. Appendix 3.A contains a proof that the error resulting from this omission depends on the pointwise mutual information (pmi) among the variables:

Proposition 1 (Underconditioning bias)**.** *Let S be a set of random variables with probability distribution $p(S)$. Let X, Y, and Z be three disjoint subsets of S. Then omitting Y from the conditioning set results in a bias in the interaction estimate I_X determined by, and linear in, the pointwise mutual information that $Y = 0$ gives about states of X:*

$$I_{X|YZ} - I_{X|Z} = \left(\prod_{i=1}^{|X|} \frac{\partial}{\partial x_i} \right) \mathrm{pmi}(X = x, Y = 0 \mid Z = 0) \tag{3.20}$$

Proposition 1 implies two positive corollaries for the practical estimation of the interactions. First, since the estimation error involves the difference between pmis, there are situations in which a variable has nonzero mutual information with the interacting variables, but omitting it from the conditioning set introduces no bias. In addition, Proposition 1 shows that excluding variables from the conditioning set for which it is hard to show dependence in the first place does not introduce a large error.

3.2.3 Estimating uncertainty and assigning significance

One of the main problems with the RBM estimation was that there is no obvious and computationally tractable way to assign a measure of uncertainty to the estimates. Using the model-free estimation outlined in this chapter, there is a very natural way to quantify the variance of the estimator with respect to the sampling distribution: bootstrap resampling. An interaction is an estimate of a property of the true population P, based on a sample S from that population. Drawing $|S|$ samples from S *with replacement* N_{bs} times generates N_{bs} new data sets $S^{(i)}$ that reflect the finite sampling variability. Each $S^{(i)}$ is called a bootstrap resample, and estimating an interaction on each of these resampled data sets yields N_{bs} bootstrap estimates for the interaction. Under the assumptions that the original data set S is a large enough and unbiased sample from the true population P, and that the samples in S are independent, the variance of the estimate across the bootstrap resamples is asymptotically (in the number of resamples N_{bs}) consistent with the sampling variance that results from finite sampling from P.

Practically, one chooses N_{bs} large enough so that the bootstrap statistics stabilise. The most important property of the interaction estimate is whether it is significantly nonzero. This is quantified by the size of the 1-sided confidence interval that does not include zero. Using the so-called percentile bootstrap procedure, this is calculated as follows: If the point estimate—*i.e.* the estimate on the original data set S—has sign s, then the fraction F of bootstrap estimates with sign $-s$ quantifies the certainty with which that interaction is nonzero. On top of this, one might also be interested in the exact value of the interaction, in which case the symmetric 95% confidence interval around the median is more informative. Figure 3.3 shows both the F-value and the size of the 95% confidence interval as a function of the number of bootstrap resamples, for N_{bs} up to 2k. It can be seen that the big fluctuations indeed happen for low N_{bs}, and that $N_{bs} = 1,000$ suffices to stabilise the uncertainty estimates. In fact, out of 2k randomly chosen pairs, no interaction lost its significance at $F \leq 0.05$ when increasing the number of resamples from 1k to 2k. One interaction gained significance after this increase, but the associated F-values were 0.0520 and 0.0495, respectively, so this does not represent a large instability. I therefore set $N_{bs} = 1,000$ throughout this thesis.

Note that even when using a Markov blanket to relax the conditioning in Equation (3.19), estimating the sample means might be impossible if there are no samples available that satisfy the condition. In this case, the expectation value, and by extension the interaction, are deemed *inestimable*. Throughout this thesis, interactions with an inestimable point estimate are ignored, but it is important to emphasise that this does not mean these interactions are zero, weak, or insignificant.

Even if an interaction is estimable on S, it might not be estimable on a particular bootstrap resample $S^{(i)}$. In that case, the bootstrap distribution can look pathological, and the uncertainty statistics it implies are no longer valid. Figure 3.4 shows four examples of distributions of bootstrap estimates. It can be seen that the distributions quickly become multimodal as u, the fraction of undefined bootstrap estimates, increases. Therefore, throughout this thesis, I will only consider interactions that are perfectly estimable, *i.e.* the interaction was estimable on all resampled data sets $S^{(i)}$ (which corresponds to $u = 0$ in Figure 3.4). This is the most conservative approach to dealing with this issue, and less conservative approaches that leave more interactions estimable are outlined in Section 6.2.

These F-values are a measure of confidence that the interactions are robustly (with respect to sample variance) nonzero, but they are emphatically not p-values. The null distribution under which to evaluate the point estimate is unknown as the bootstrap distribution only describes the variability around the point estimate, not around the value zero. To construct a null distribution, one could subtract the point estimate from the bootstrap distribution to get a distribution around zero. However, this is only a valid approximation to the null distribution under the assumption that the distribution of the estimator is independent of the population parameters. In Section 6.2.1.a, I explicitly calculated the asymptotic variance of the estimator which shows that this is not the case. In other words: the MFI estimator is not an ancillary/pivotal statistic. This means that the bootstrap distribution cannot be moved to zero to construct a null distribution, and that assigning p-values to the estimates is therefore not possible. This also makes multiple testing or false discovery rate correction on the F-values meaningless. For this reason, I decided to quantify the confidence level that an interaction is nonzero by the F-value directly.

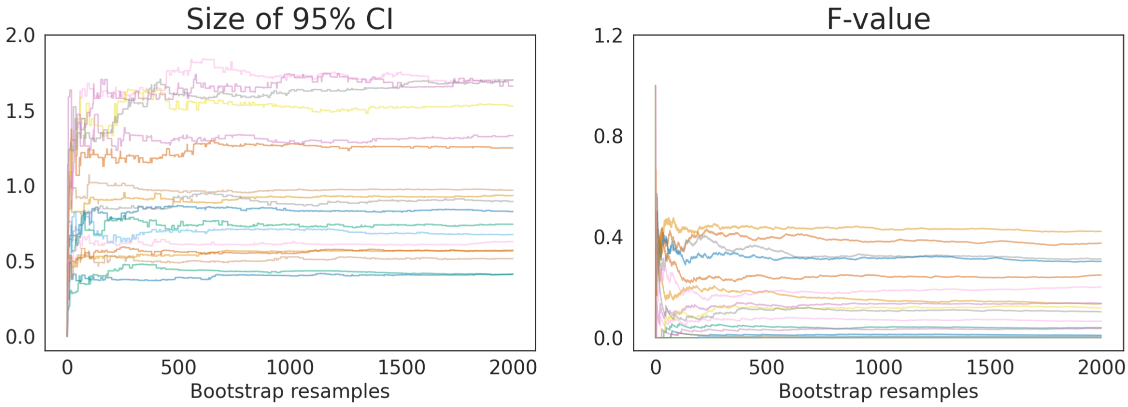


Figure 3.3: The F-value and the size of the 95% confidence interval (CI) both stabilised at around 1,000 bootstrap resamples. Shown here for 20 randomly chosen 2-point interactions in a data set that contained 20,000 neurons and astrocytes from the developmental scRNA-seq data set that is introduced in Section 4.2.6.

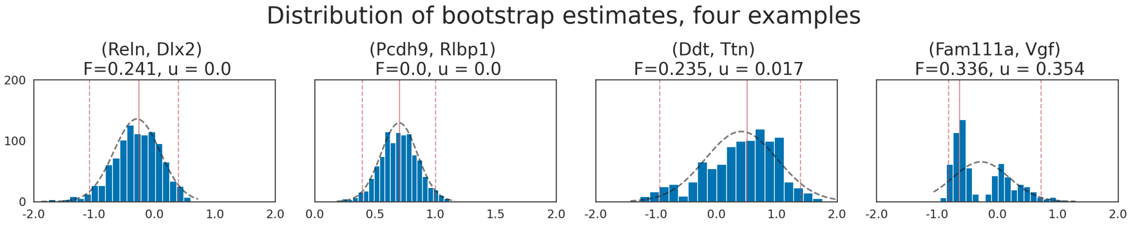


Figure 3.4: Four examples of bootstrap distributions of pairwise MFIs in the same data set as Figure 3.3. When an interaction was not perfectly estimable—*i.e.* when some of the resampled estimates were undefined or diverge—the bootstrap distribution became pathological or multimodal. From left to right: a perfectly estimable non-significant interaction (between *Reln* and *Dlx2*), a perfectly estimable significant interaction (between *Pcdh9* and *Rlbp1*), a marginally estimable interaction (between *Ddt* and *Ttn*), and a non-estimable interaction (between *Fam111a* and *Vgf*). F is the significance F-value, u is the fraction of undefined resamples. Dashed red lines denote the 95% CI, and the solid red line denotes the original point estimate on S. In dashed grey: a normal distribution of matched mean and variance.

3.3 Results

Section 3.3.1 shows a theoretical result that explicitly links the MFIs to information theory through something called Möbius inversions. Then, in **Section 3.3.2**, I analytically calculated higher-order interactions in logic gates and show how they are related to the synergistic logic represented by the gate. I then moved on to more complex causal dynamics and calculated MFIs on simulated data in **Section 3.3.3**.

3.3.1 MFIs are Möbius inversions of surprisal

> One cannot invent the structure of an object. The most we can do is to patiently bring it to the light of day, with humility—in making it known, it is discovered.
>
> Alexander Grothendieck [98]

This section explicitly links the MFIs to information theory. To do so, Section 3.3.1.a first recasts information theory in terms of Möbius inversions. Section 3.3.1.b then defines MFIs in terms of Möbius inversions and contains a proof that this new definition is equivalent to Definition 5. Finally, Section 3.3.1.c uses the Möbius inversion formalism to define quantities dual to mutual information and the MFIs, which both turn out to be well-defined and useful objects.

3.3.1.a Mutual information as a Möbius inversion

Consider the definition of mutual information, and its higher-order generalisation in terms of the entropy function H:

- Mutual information:

$$\begin{aligned} MI(X,Y) &= H(X) - H(X \mid Y) && (3.21)\\ &= H(X) + H(Y) - H(X,Y) && (3.22)\end{aligned}$$

- The generalisation of mutual information that describes higher-order dependencies between variables goes by many names: *multiple mutual information, co-information*, or *interaction information*, and is defined on three variables as follows:

$$\begin{aligned} MI(X,Y,Z) =& MI(X,Y) - MI(X,Y \mid Z) && (3.23)\\ =& H(X) + H(Y) + H(Z) \\ &- H(X,Y) - H(X,Z) - H(Y,Z) \\ &+ H(X,Y,Z) && (3.24)\end{aligned}$$

I will refer to both these quantities simply as mutual information (MI), and combine them into one definition in terms of a Möbius inversion, which will be defined below.

Note that each MI-based quantity can be written as a specific sum of marginal entropies of subsets of the set of variables. There is structure among these subsets: given a finite

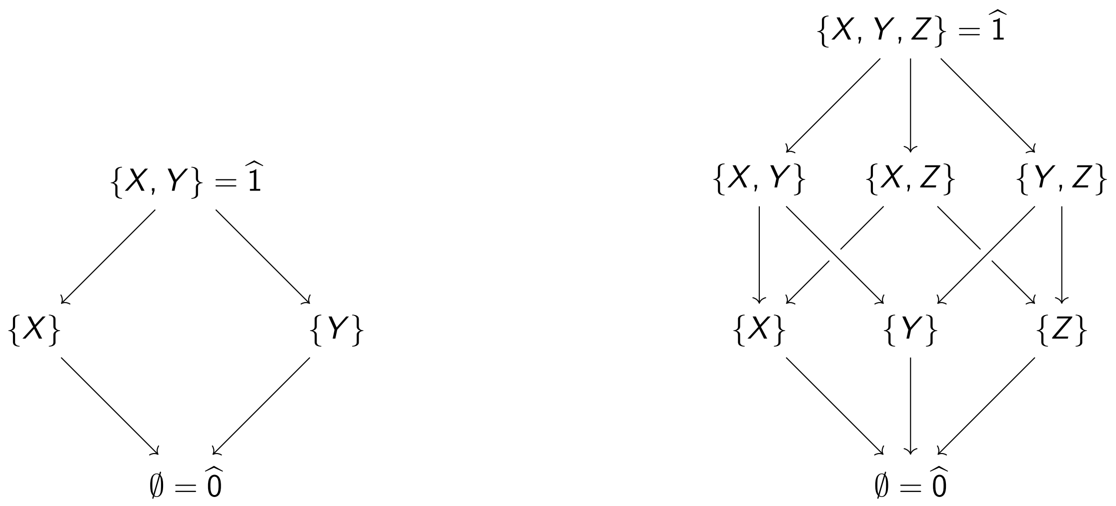


Figure 3.5: The lattices associated to $\mathcal{P}(\{X, Y\})$ (left) and $\mathcal{P}(\{X, Y, Z\})$ (right), ordered by inclusion. An arrow $b \to a$ indicates $a < b$. The 'top' and 'bottom' elements are also denoted by $\widehat{1}$ and $\widehat{0}$, respectively.

set of variables S, its powerset $\mathcal{P}(S)$ can be given a partial ordering as follows:

$$a \leq b \iff a \subseteq b \quad \forall\, a, b \in \mathcal{P}(S) \tag{3.25}$$

This poset $P = (\mathcal{P}(S), \subseteq)$ is called a Boolean algebra, and since each pair of sets has a unique supremum (their union) and infimum (their intersection), it is a lattice. This lattice structure is visualised for two and three variables in Figure 3.5. In general, the lattice of an n-variable Boolean algebra forms an n-cube. Moreover, Boolean algebras are *locally finite*[2], so for any Boolean algebra P one can define the *incidence algebra* of functions from nonempty intervals on P to $\mathbb{R}$. Important elements of this algebra are *e.g.* the identity element $\delta(a, b) = \delta_{ab}$ (the Kronecker delta), and the constant function $\zeta(a, b) = 1$. Of particular interest is the algebraic inverse of the constant zeta-function: the Möbius function $\mu_P : P \times P \to \mathbb{R}$, defined as

$$\mu_P(x, y) = \begin{cases} 1 & \text{if } x = y \\ -\sum\limits_{z : x \leq z < y} \mu_P(x, z) & \text{if } x < y \\ 0 & \text{otherwise} \end{cases} \tag{3.26}$$

On a powerset ordered by inclusion, the Möbius function takes the simple form $\mu(x, y) = (-1)^{|x|-|y|}$ [254, 224]. This definition allows the mutual information among a set of variables τ to be written as [25, 86]:

$$\begin{aligned} MI(\tau) &= (-1)^{|\tau|-1} \sum_{\eta \leq \tau} \mu_P(\eta, \tau) H(\eta) \qquad (3.27) \\ &= \sum_{\eta \leq \tau} (-1)^{|\eta|+1} H(\eta) \qquad (3.28) \end{aligned}$$

[2] A poset P is locally finite if for any two elements a and b, the set $[a, b] = \{x : \; a \leq x \leq b\}$, called a *closed interval*, is finite.

Where P is the Boolean algebra with $\tau = \widehat{1}$, and $H(\eta)$ is the marginal entropy of the set of variables η. This indeed coincides with Equation (3.22) for $\tau = \{X, Y\}$ and Equation (3.24) for $\tau = \{X, Y, Z\}$. Equation (3.27) is a convolution known as a Möbius inversion:

Definition 6 (Möbius inversion over a poset, Rota 1964 [224])**.** *Let P be a locally finite poset $(S, \leq)$. Let $\mu_P : P \times P \to \mathbb{R}$ be the Möbius function from Equation (3.26). Let $g : P \to \mathbb{R}$ be a function on P. Then the function*

$$f(y) = \sum_{x \leq y} \mu_P(x, y) g(x) \tag{3.29}$$

is called the Möbius inversion of g on P. Furthermore, this equation can be inverted to yield

$$f(y) = \sum_{x \leq y} \mu_P(x, y) g(x) \iff g(y) = \sum_{x \leq y} f(x) \tag{3.30}$$

The Möbius inversion is a generalisation of the discrete derivative to locally finite posets. If $P = (\mathbb{N}, \leq)$, Equation (3.30) is just a discrete version of the fundamental theorem of calculus [254]. Equation (3.30) also implies that the joint entropy can be expressed as a sum over mutual information:

$$H(\tau) = (-1)^{|\tau|-1} \sum_{\eta \leq \tau} MI(\eta) \tag{3.31}$$

For example, in the case of three variables:

$$H(X, Y, Z) = MI(X, Y, Z) + MI(X, Y) + MI(X, Z) + MI(Y, Z) + H(X) + H(Y) + H(Z) \tag{3.32}$$

Instead of starting with entropy, one could also start with a quantity known as surprisal, or self-information, defined as the negative log-probability of a certain state:

$$S(X = x) = -\log p(X = x) \tag{3.33}$$

Surprisal plays an important role in information theory, and indeed, the expected surprisal across all possible realisations $X = x$ is the entropy of the variables X:

$$\mathbb{E}_X[S(X = x)] = H(X) \tag{3.34}$$

As a shorthand for the marginal surprisal of a realisation $X = x$, summed over Y, let us write

$$\log p(x; Y) := \sum_{y} \log p(x, y) \tag{3.35}$$

With this, consider the Möbius inversion of the marginal surprisal over the lattice P:

$$\text{pmi}(T = \tau) := (-1)^{|\tau|} \sum_{\eta \leq \tau} \mu_P(\eta, \tau) \log p(\eta; \tau \setminus \eta) \tag{3.36}$$

This is a generalised version of the pointwise mutual information, usually defined on just two variables:

$$\begin{aligned}\mathrm{pmi}(X=x, Y=y) &= \log(x, y; \emptyset) - \log(x; Y) - \log(y; X) + \log(\emptyset; X, Y) &\quad (3.37)\\ &= \log \frac{p(x,y)}{p(x)p(y)} &\quad (3.38)\end{aligned}$$

3.3.1.b MFIs as a Möbius inversion

With mutual information defined in terms of Möbius inversions, the same can be done for the model-free interactions. Consider, again, the negative surprisal of a particular state. On Boolean variables, a state is just a partition of the variables into two sets: one where the variables are set to 1, and one where they are set to 0. That means that the surprisal of observing a particular state of Z variables is fully specified by which variables $X \subseteq Z$ are set to 1, keeping all other variables $Z \setminus X$ at 0. This can be written as:

$$S_{X;Z} := \log p(X=1, Z \setminus X = 0) \qquad (3.39)$$

Definition 7 (Interactions as Möbius inversions)**.** *Let p be a probability distribution over a set T of random variables. Let $P = (\mathcal{P}(\tau), \subseteq)$, the powerset of a set $\tau \subseteq T$ ordered by inclusion. Then the interaction $I(\tau; T)$ among variables τ is given by*

$$\begin{aligned}I(\tau; T) &:= \sum_{\eta \leq \tau} \mu_P(\eta, \tau) S_{\eta;T} &\quad (3.40)\\ &= \sum_{\eta \leq \tau} (-1)^{|\eta|-|\tau|} \log p(\eta = 1, T \setminus \eta = 0) &\quad (3.41)\end{aligned}$$

For example, when τ contains a single variable $X \subseteq T$, then

$$\begin{aligned}I(\{X\}; T) &= \mu_P(\{X\}, \{X\}) S_{\{X\};T} + \mu_P(\emptyset, \{X\}) S_{\emptyset;T} &\quad (3.42)\\ &= \log \frac{p(X=1, T \setminus X = 0)}{p(X=0, T \setminus X = 0)} &\quad (3.43)\end{aligned}$$

Which coincides with the 1-point interaction in Equation (3.9). Similarly, when τ contains two variables $\tau = \{X, Y\} \subseteq T$, then

$$\begin{aligned}I(\{X,Y\}; T) = \mu_P(\{X,Y\}, \{X,Y\}) S_{\{X,Y\};T} + \mu_P(\{X\}, \{X,Y\}) S_{\{X\};T} &\quad (3.44)\\ +\mu_P(\{Y\}, \{X,Y\}) S_{\{Y\};T} + \mu_P(\emptyset, \{X,Y\}) S_{\emptyset;T} &\\ = \log \frac{p(X=1, Y=1, T \setminus \{X,Y\} = 0) p(X=0, Y=0, T \setminus \{X,Y\} = 0)}{p(X=1, Y=0, T \setminus \{X,Y\} = 0) p(X=0, Y=1, T \setminus \{X,Y\} = 0)} &\quad (3.45)\end{aligned}$$

Which coincides with the 2-point interaction in Equation (3.11). In fact, this pattern holds in general:

Theorem 2 (Equivalence of interactions)**.** *The interaction $I(\tau, T)$ from Definition 7 is the same as the model-free interaction I_τ from Definition 5. That is, for any set of variables $\tau \subseteq T$*

$$I(\tau, T) = I_\tau \tag{3.46}$$

Proof. I intend to prove the following:

$$\sum_{\eta \leq \tau} (-1)^{|\eta|-|\tau|} \log p(\eta = 1, T \setminus \eta = 0) = \frac{\partial^n \log p(T)}{\partial \tau_1 \dots \partial \tau_n}\bigg|_{\underline{T}=0} \tag{3.47}$$

where $\tau = \{\tau_1, \dots \tau_n\}$. Both sides of this equation are sums of $\pm \log p(s)$, where s is some binary string, so the same strings should appear with the same sign.

First, note that the Boolean algebra of sets ordered by inclusion (as in Figure 3.5), is equivalent to the poset of binary strings where for any two strings a and b, $a \leq b \iff a \wedge b = a$. The equivalence follows immediately upon setting each element $a \in \mathcal{P}(S)$ to the string where $a = 1$ and $S \setminus a = 0$. This map is one-to-one and monotonic with respect to the partial order as $A \subseteq B \iff A \cap B = A$. That means that Definition 7 can be written as a Möbius inversion on the lattice of Boolean strings $S = (\mathbb{B}^{|\tau|}, \leq)$ (shown for the 3-variable case on the left side of figure 3.6):

$$I(\tau; T) = \sum_{s \leq \hat{1}_S} \mu_S(s, \hat{1}_S) \log p(\tau = s, T \setminus \tau = 0) \tag{3.48}$$

Note that for any pair (α, τ) where $\alpha \subseteq \tau$, with respective string representations $(s, t) \in \mathbb{B}^{|\tau|} \times \mathbb{B}^{|\tau|}$, the following holds:

$$|\tau| - |\alpha| = \sum_i (t \wedge \neg s)_i \tag{3.49}$$

Since τ corresponds to a string of ones, $I(\tau; T)$ can be written as:

$$I(\tau; T) = \sum_{s \leq \hat{1}_S} (-1)^{\sum \neg s} \log p(\tau = s, T \setminus \tau = 0) \tag{3.50}$$

To see that this is exactly the Boolean derivative from Definition 5, define a map

$$e^{(n)}_{i,s} : \mathcal{F}_{\mathbb{B}^n} \rightarrow \mathcal{F}_{\mathbb{B}^{n-1}} \tag{3.51}$$

where $\mathcal{F}_{\mathbb{B}^n}$ is the set of functions from n Boolean variables to $\mathbb{R}$. This map is defined as

$$e^{(n)}_{i,s} : f(X_0, \dots X_i, \dots X_n) \mapsto f(X_0, \dots X_i = s, \dots X_n) \tag{3.52}$$

With this map, the Boolean derivative of a function $f(X_0, \dots, X_n)$ can be written as

$$\frac{\partial}{\partial X_i} f(X) = (e^{(n)}_{i,1} - e^{(n)}_{i,0}) f(X) \tag{3.53}$$

$$= f(X_1, \dots, X_i = 1, \dots, X_n) - f(X_1, \dots, X_i = 0, \dots, X_n) \tag{3.54}$$

Such that the derivative w.r.t. a set S of m variables becomes function composition:

$$\left(\prod_{i=0}^{m} \frac{\partial}{\partial X_{S_i}}\right) f(X) = \left(\bigcirc_{i=0}^{m} (e_{S_i,1}^{(n-i)} - e_{S_i,0}^{(n-i)})\right) f(X) \tag{3.55}$$

From this, it is clear that a term $f(s)$ appears with a minus sign iff $e_{i,0}^{(n)}$ has been applied an odd number of times. Therefore, terms where s contains an odd number of 0s get a minus sign. This can be summarised as:

$$\left(\prod_{i=0}^{m} \frac{\partial}{\partial X_{S_i}}\right) f(X) = \sum_{s \in \mathbb{B}^n} (-1)^{\sum \neg s} f(X_S = s, X \setminus X_S) \tag{3.56}$$

and therefore

$$I_\tau = \sum_{s \in \mathbb{B}^n} (-1)^{\sum \neg s} \log(\tau = s, T \setminus \tau = 0) \tag{3.57}$$

Noting that the sums $\sum_{s \leq \hat{1}_S}$ and $\sum_{s \in \mathbb{B}^n}$ contain exactly the same terms equates Equations (3.57) and (3.50), and completes the proof.

□

Note that the structure of the lattice S reveals some structure in the interactions that was already noted in [24]. The rightmost lattice in Figure 3.6 shows the 3-variable lattice again, this time with two shaded regions. The green region corresponds to the 2-point interaction between the first two variables. The red region is a similar interaction between the first two variables, but in the context of the third variable fixed to 1, instead of 0. This shows the interpretation of a 3-point interaction as the difference in two 2-point interactions: $I_{XYZ} = I_{XY}|_{Z=1} - I_{XY}$. The symmetry of the hypercube shows the three different but equivalent choices for which variable to set to 1. Treating the Boolean algebra as a die where the sides facing up are ⚀, ⚁, and ⚂, the 3-point interaction is given by the difference between antipodal faces.

$$I_{XYZ} = \text{⚀} - \text{⚅} = \text{⚁} - \text{⚄} = \text{⚂} - \text{⚃} \tag{3.58}$$

As before, Definition 7 can be inverted to express the surprise of observing a state with all 1s in terms of interactions:

$$\log p(\tau = 1, T \setminus \tau = 0) = \sum_{\eta \leq \tau} I(\eta, T) \tag{3.59}$$

For example, in the case where $T = \{X, Y, Z\}$ and $\tau = \{X, Y\}$

$$S(1,1,0) = -\log p(1,1,0) = -I_{XY} - I_X - I_Y - I_\emptyset \tag{3.60}$$

which illustrates that when X and Y tend to be off ($I_X < 0$ and $I_Y < 0$), and X and Y tend to be different ($I_{XY} < 0$), then observing the state (1, 1, 0) is very surprising, as should be expected.

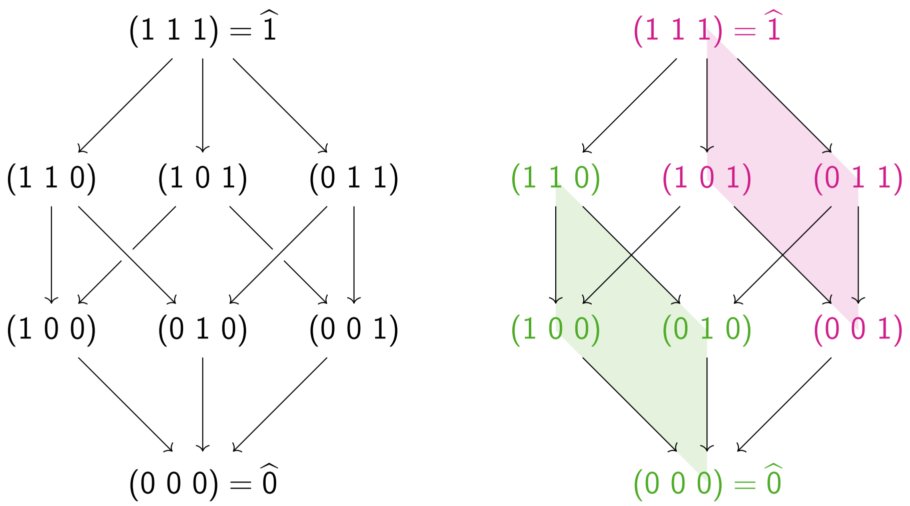


Figure 3.6:
Left: The lattice associated to $\mathcal{P}(\{X, Y, Z\})$ ordered by inclusion, as binary strings. Equivalently: the lattice of binary strings, where for any two strings a and b, $a \leq b \iff a \wedge b = a$.
Right: Two regions are shaded, corresponding to the decomposition of the 3-point interaction into two 2-point interactions.

Categorical interactions Taking the definition of interactions as the Möbius inversion of surprisal seriously, one might ask what happens when instead of using a Boolean algebra, surprisal is inverted over a different lattice. One example of a different lattice is shown in Figure 3.7. It corresponds to two variables that can take three values—0, 1, or 2—where states are ordered by $a \leq b \iff \forall i : a_i \leq b_i$. Calculating interactions on this lattice requires the value of Möbius functions of the type $\mu(s, 22)$. It can be readily verified that most Möbius functions like this are zero, except for $\mu(22, 22) = \mu(11, 22) = 1$, and $\mu(21, 22) = \mu(12, 22) = -1$, which gives exactly the terms in the interactions between two categorical variables changing from $1 \rightarrow 2$, as defined in [24]. Calculating interactions on different sublattices with $\widehat{1} = (21), (12)$, or (11) gives the other categorical interactions. The transitivity property of the interactions, i.e. $I(X : 0 \rightarrow 2, Y : 0 \rightarrow 1) = I(X : 0 \rightarrow 1, Y : 0 \rightarrow 1) + I(X : 1 \rightarrow 2, Y : 0 \rightarrow 1)$, follows immediately from the structure of the lattice in Figure 3.7, and the alternating signs of the Möbius functions on a Boolean algebra.

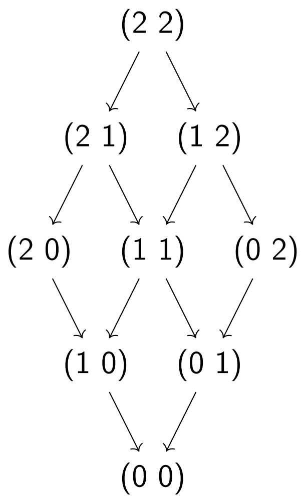


Figure 3.7: The lattice of two variables that can take three values, ordered by $a \leq b \iff \forall i : a_i \leq b_i$.

3.3.1.c Information and interactions on dual lattices

Lattices have the property that the set with the reverse order is still a lattice. That is, if $\mathcal{L} = (S, \leq)$ is a lattice, then $\mathcal{L}^{\text{op}} = (S, \preceq)$, where $\forall a, b \in S : a \preceq b \iff a \geq b$, is also a lattice. This raises the question: what corresponds to mutual information and interaction on these dual lattices[3]?

Dual information The quantity dual to mutual information, denoted by MI^*, can be calculated by first noting that the dual to a Boolean algebra is another Boolean algebra—so that it is still true that $\mu(x, y) = (-1)^{|x|-|y|}$—and then simply replacing P by P^{op} in Equation (3.27):

$$MI^*(\tau) = \sum_{\eta \preceq \tau} (-1)^{|\eta|+1} H(\eta) \tag{3.61}$$

The dual mutual information of $\tau = \widehat{1}_{P^{\text{op}}} = \emptyset$ is just $MI^*(\emptyset) = MI(\widehat{1}_P)$, the mutual information among all variables. However, the dual mutual information of a singleton set X is:

$$\begin{aligned} MI^*(X) &= MI(\widehat{1}_P) - MI(\widehat{1}_P \setminus X) && (3.62)\\ &= \Delta(X; \widehat{1}_P \setminus X) && (3.63) \end{aligned}$$

where Δ is known as the conditional, or differential mutual information [87]. It describes the change in mutual information when leaving out X, and has been used to describe information structures in genetics [86]. On the Boolean algebra of three variables $\{X, Y, Z\}$, the dual mutual information of X can be written out as:

[3] Recognising that a poset $\mathcal{L} = (S, \leq_{\mathcal{L}})$ is a category $\mathcal{C}$ with objects S and a morphism $f : A \to B$ iff $B \leq_{\mathcal{L}} A$, reversing the order defines the *opposite* category $\mathcal{C}^{\text{op}}$, and thus dual objects.

$$MI^*(X) = \mu(\{X\}, \{X\})H(X) + \mu(\{X,Y\}, \{X\})H(X,Y)+ \\ \mu(\{X,Z\}, \{X\})H(X,Z) + \mu(\{X,Y,Z\}, \{X\})H(X,Y,Z) \tag{3.64}$$

$$= H(X) - H(X,Y) - H(X,Z) + H(X,Y,Z) \tag{3.65}$$

Since Δ is the dual of mutual information, it should arguably be called the mutual co-information, but the term co-information is unfortunately already used to refer to normal higher-order mutual information.

Outeractions To find the quantity dual to the interactions, start from Equation (3.48) and construct $S^{\text{op}} = (\mathbb{B}^{|\tau|}, \preceq)$, dual to the lattice of binary strings $S = (\mathbb{B}^{|\tau|}, \leq)$. A dual interaction of variables $\tau \subseteq T$ will be denoted $I^*(\tau; T)$, and is defined as follows:

$$I^*(\tau; T) := \sum_{s \preceq \widehat{1}_{S^{\text{op}}}} \mu_{S^{\text{op}}}(s, \widehat{1}_{S^{\text{op}}}) \log p(\tau = s, T \setminus \tau = 0) \tag{3.66}$$

Again, when $\tau = \widehat{1}_{S^{\text{op}}} = \widehat{0}_S = \emptyset$, this is just $(-1)^{|\tau|} I(\widehat{1}_S)$, but the dual interaction of a singleton set X is:

$$I^*(X; T) = (-1)^{|\widehat{1}_S|-1}\Big(I(\widehat{1}_S; T) + I(\widehat{1}_S \setminus X; T)\Big) \tag{3.67}$$

For example, on the three variable lattice in Figure 3.6, the dual interaction of X is

$$I^*(X; T) = I(X,Y,Z; T) + I(Y,Z; T) \tag{3.68}$$

Writing p_{ijk} for $p(X = i, Y = j, Z = k \mid T \setminus \{X,Y,Z\} = 0)$, we see that this is equal to:

$$I^*(X; T) = \log \frac{p_{111}p_{100}}{p_{101}p_{110}} \tag{3.69}$$

which is similar to the 2-point interaction I_{YZ} defined in Equation (3.11), but conditioned on $X = 1$ instead of 0. Dual interactions should probably be called co-interactions, but to avoid confusion with the term co-information I will instead refer to the dual interactions as outeractions. Outeractions are just interactions, conditioned on certain variables being 1 instead of 0. This makes them no longer equal to the Ising interactions between Boolean variables, but there are situations in which an interaction is more interesting in the context with $Z = 1$ instead of $Z = 0$, for example if Z is always 1 in all situations of interest.

Summary

- *Mutual information is the Möbius inversion of marginal entropy on the lattice of subsets ordered by inclusion.*
- *Pointwise mutual information is the Möbius inversion of marginal surprisal on the lattice of subsets ordered by inclusion.*

- *Differential (or conditional) mutual information is the Möbius inversion of marginal entropy on the dual lattice.*
- *Model-free interactions are the Möbius inversion of surprisal on the lattice of subsets ordered by inclusion.*
- *Model-free outeractions are the Möbius inversion of surprisal on the dual lattice.*
- *The outeraction of a variable X is the interaction among its complement, where X is set to 1 instead of 0.*

To summarise these relationships diagrammatically, note that surprisals form a vector space as follows. Let $\mathcal{P}(T)$ be the powerset of a set of variables T, and let $|\mathcal{P}(T)| = 2^{|T|} := n$. This forms the lattice $P = (\mathcal{P}(T), \subseteq)$ ordered by inclusion, and a linear extension[4] of P induces a topological ordering on $\mathcal{P}(T)$, indexed by i as $\mathcal{P}(T) = \cup_{i=0}^{n} t_i$ (that is, the set of t_i forms a cover of $\mathcal{P}(T)$). Let $\mathcal{S}$ be the set of linear combinations of surprisals of subsets of T:

$$\mathcal{S} = \left\{ \sum_{i=0}^{n} a_i \log p(t_i) \mid a_i \in \mathbb{R} \right\} \tag{3.70}$$

This set is given a vector space structure over $\mathbb{R}$ by the usual scalar multiplication and addition. Note that the set

$$\mathcal{B} = \{\log p(t) \mid t \in \mathcal{P}(T)\} \tag{3.71}$$

forms a basis for this vector space, since $\sum_i \alpha_i \log p(t_i) = 0$ has no non-trivial solutions[5], and $\text{span}(\mathcal{B}) = \mathcal{S}$. To define a map from $\mathcal{S} \to \mathbb{R}$, we only need to specify its action on $\mathcal{B}$, and extend the definition linearly. That means we can fully define the map $eval_T : \mathcal{S} \to \mathbb{R}$ by specifying:

$$eval_T : \log p(R = r) \mapsto \log p(R = 1, T \setminus R = 0) \tag{3.72}$$

Similarly, define the expectation map $\mathbb{E} : \mathcal{S} \to \mathbb{R}$ as

$$\mathbb{E} : \log p(R = r) \mapsto \sum_{r} p(R = r) \log(R = r) \tag{3.73}$$

which outputs the expected surprise over all realisations $R = r$. Finally, note that the Möbius inversion over a poset P is an endomorphism of the set $\mathcal{F}_P$ of functions over P, defined as

$$M_P : \mathcal{F}_P \to \mathcal{F}_P \tag{3.74}$$

$$M_P : f(y) \mapsto \sum_{x \leq y} \mu(x, y) f(x) \tag{3.75}$$

[4] Interestingly, the existence proof of this linear extension, known as the Szpilrajn extension theorem, depends on the axiom of choice.

[5] Only when two variables a and b are independent do we have linear dependencies in $\mathcal{B}$, as then $\log p(a, b) = \log p(a) + \log p(b)$.

Together, these three maps make the following squares commute (here shown on individual elements):

$$\begin{tikzcd}
MI^*(R) = MI(T \setminus R \mid R) & -H(R) \arrow[l, mapsto, "M_{Pop}"'] \arrow[r, mapsto, "M_P"] & MI(R) \\
\mathrm{pmi}^*(R = r) \arrow[u, mapsto, "\mathbb{E}"'] \arrow[d, mapsto, "eval_T"] & \log p(R = r) \arrow[l, mapsto, "M_{Pop}"'] \arrow[r, mapsto, "M_P"] \arrow[u, mapsto, "\mathbb{E}"'] \arrow[d, mapsto, "eval_T"] & \mathrm{pmi}(R = r) \arrow[u, mapsto, "\mathbb{E}"'] \arrow[d, mapsto, "eval_T"] \\
I^*(R; T) & \log p(R = 1; T = 0) \arrow[l, mapsto, "M_{Pop}"'] \arrow[r, mapsto, "M_P"] & I(R; T)
\end{tikzcd}$$

For the case where $T = \{X, Y, Z\}$ and $R = \{X, Y\}$, this explicitly amounts to:

$$\begin{tikzcd}
\begin{array}{c} \sum_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} p(x,y,z) \log p(x,y,z) \\ -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x,y) \log p(x,y) \end{array} & \sum_{(x,y)\in X\times Y} p(x,y) \log p(x,y) \arrow[l, mapsto, "M_{Pop}"'] \arrow[r, mapsto, "M_P"] & \begin{array}{c} \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x,y) \log p(x,y) \\ -\sum_{x\in\mathcal{X}} p(x) \log p(x) \\ -\sum_{y\in\mathcal{Y}} p(y) \log p(y) \end{array} \\
\log \frac{p(x,y,z)}{p(x,y)} \arrow[u, mapsto, "\mathbb{E}"'] \arrow[d, mapsto, "eval_T"] & \log p(x,y) \arrow[l, mapsto, "M_{Pop}"'] \arrow[r, mapsto, "M_P"] \arrow[u, mapsto, "\mathbb{E}"'] \arrow[d, mapsto, "eval_T"] & \log \frac{p(x,y)p(\emptyset)}{p(x)p(y)} \arrow[u, mapsto, "\mathbb{E}"'] \arrow[d, mapsto, "eval_T"] \\
\log \frac{p(1,1,1)}{p(1,1,0)} & \log p(1,1,0) \arrow[l, mapsto, "M_{Pop}"'] \arrow[r, mapsto, "M_P"] & \log \frac{p(1,1,0)p(0,0,0)}{p(1,0,0)p(0,1,0)}
\end{tikzcd}$$

3.3.2 Interactions quantify and distinguish synergistic logic

In this section, I will show that a nonzero 3-point interaction I_{ABC} on a causal collider structure $A \to C \leftarrow B$ can be interpreted as a logic gate. In fact, I will show that the MFIs are better at distinguishing logic gates than information theoretic quantities.

A positive 3-point interaction implies that the numerator in Equation (3.14) is larger than the denominator. The sufficient but not necessary assumption that each term in the numerator is larger than each term in the denominator results in the following truth table as $I_{ABC} \to +\infty$:

A	B	C
0	0	1
0	1	0
1	0	0
1	1	1

This is the truth table of an XNOR gate. Let $p_{\mathcal{G}}$ be the probability of each of the four states in the truth table for a gate $\mathcal{G}$, and let $\epsilon_{\mathcal{G}}$ be the probability of all other states. Then the 3-point interaction of an XNOR gate can be written as:

$$I_{ABC}^{\text{XNOR}} = \log \frac{p_{\text{XNOR}}^4}{\epsilon_{\text{XNOR}}^4} \tag{3.76}$$

Similarly, from the truth tables of AND and OR gates:

$$I_{ABC}^{\text{AND}} = \log \frac{\epsilon_{\text{AND}}\ p_{\text{AND}}^3}{\epsilon_{\text{AND}}^3 p_{\text{AND}}} \tag{3.77}$$

$$I_{ABC}^{\text{OR}} = \log \frac{\epsilon_{\text{OR}}^3 p_{\text{OR}}}{\epsilon_{\text{OR}}\ p_{\text{OR}}^3} \tag{3.78}$$

In the case of equally noisy gates, *i.e.* $p_{\mathcal{G}} = p$ and $\epsilon_{\mathcal{G}} = \epsilon$, the associated 3-point interactions can be directly compared. Note that when a gate has a 3-point interaction I, its logical negation a 3-point interaction $-I$. This determines the 3-point interactions of all $2^3 = 6$ possible 2-input logical gates, summarised in Table 3.1. The two gates with the strongest absolute interactions, XNOR and XOR, are also the only two gates that are purely synergistic: knowing just one of the two inputs gives you no information about the output. This relationship to synergy holds for 3-input gates as well. The 3-input gate with the strongest 4-point interaction has the following truth table:

A	B	C	D
0	0	0	0
0	0	1	1
0	1	0	1
1	0	0	1
0	1	1	0
1	0	1	0
1	1	0	0
1	1	1	1

It is a 3-input XOR gate, i.e. $D = (A + B + C) \mod 2$, and is again maximally synergistic since observing only 2 of the 3 inputs gives zero bits of information on the output. Setting this maximum 4-point interaction to I, the 3-input OR and AND gates get 4-point interaction $I/4$, so the hierarchies of interaction and synergy still match.

As can be seen in Table 3.1, the 3-point interactions separate most 2-input logic gates by sign or value, leaving only AND $\sim$ NOR and OR $\sim$ NAND. Mutual information has less resolving power. Assuming a uniform distribution over all 4 allowed states from a

$\mathcal{G}$	$I^{\mathcal{G}}_{ABC}$
XNOR	I
XOR	$-I$
AND	$\frac{1}{2}I$
OR	$-\frac{1}{2}I$
NAND	$-\frac{1}{2}I$
NOR	$\frac{1}{2}I$

Table 3.1: The 3-point interactions for all 2-input logic gates at equal noise level are related through $I = 4\log\frac{p}{\epsilon}$, and degenerate in AND $\sim$ NOR and OR $\sim$ NAND.

$\mathcal{G}$	$H(A)$ $=H(B)$	$H(C)$	$H(A,B)$	$H(A,C)$ $=H(B,C)$	$H(A,B,C)$
XNOR	1	1	2	2	2
XOR	1	1	2	2	2
AND	1	$\log\frac{3^{3/4}}{4}$	2	$\frac{3}{2}$	2
OR	1	$\log\frac{3^{3/4}}{4}$	2	$\frac{3}{2}$	2
NAND	1	$\log\frac{3^{3/4}}{4}$	2	$\frac{3}{2}$	2
NOR	1	$\log\frac{3^{3/4}}{4}$	2	$\frac{3}{2}$	2

Table 3.2: The marginal entropies of variables in a logic gate are degenerate in XOR $\sim$ XNOR and AND $\sim$ OR $\sim$ NAND $\sim$ NOR.

gate's truth table, a brief calculation yields:

$$MI^{\text{OR}}(A,B,C) = MI^{\text{AND}}(A,B,C) = MI^{\text{NOR}}(A,B,C) = MI^{\text{NAND}}(A,B,C) \tag{3.79}$$
$$= -\log\left(\frac{3^{3/4}}{4}\right) - 1 \approx -0.189$$

$$MI^{\text{XOR}}(A,B,C) = MI^{\text{XNOR}}(A,B,C) = -1 \tag{3.80}$$

That is, mutual information resolves strictly fewer logical gates by value, and none by sign. In fact, all entropy-based quantities necessarily inherit the degeneracy summarised in Table 3.2.

The outeractions $I^{*\mathcal{G}}_C = I^{\mathcal{G}}_{ABC} + I^{\mathcal{G}}_{AB}$ contain the same degeneracy as the 3-point interactions. However, let us use the same sign-convention as differential mutual information and define a new quantity $J^{*\mathcal{G}}_A = I^{\mathcal{G}}_{ABC} - I^{\mathcal{G}}_{BC}$. This quantity assigns a different value to each logic gate $\mathcal{G}$. The symmetric quantity $\overline{J}^{*\mathcal{G}} = J^{*\mathcal{G}}_A J^{*\mathcal{G}}_B J^{*\mathcal{G}}_C$, the interaction analogous to the symmetric deltas from [87], inherits the perfect resolution from $J^{*\mathcal{G}}_A$. This is summarised in Table 3.3. The J-outeractions thus uniquely assign a value to each gate, proportional to the synergy of its logic. The hierarchy is $J^{*\text{XNOR}}_A > J^{*\text{NOR}}_A > J^{*\text{AND}}_A$, mirrored for their logical complement. XNOR is indeed the most synergistic, and NOR is more synergistic than AND with respect to observing a 0 in one of the inputs: in a NOR gate, a 0 in the input gives no information on the output, while it completely fixes

$\mathcal{G}$	MI_{ABC}	$I^{\mathcal{G}}_{ABC}$	$I^{*\mathcal{G}}_{A}$	$J^{*\mathcal{G}}_{A}$	$J^{*\mathcal{G}}_{C}$	$\overline{J}^{*\mathcal{G}}$
XNOR	-1	I	$\frac{1}{2}I$	$\frac{3}{2}I$	$\frac{3}{2}I$	$\frac{27}{8}I^3$
XOR	-1	$-I$	$-\frac{1}{2}I$	$-\frac{3}{2}I$	$-\frac{3}{2}I$	$-\frac{27}{8}I^3$
AND	-0.189	$\frac{1}{2}I$	$\frac{1}{2}I$	$\frac{1}{2}I$	$\frac{3}{4}I$	$\frac{3}{16}I^3$
OR	-0.189	$-\frac{1}{2}I$	0	$-I$	$-\frac{3}{4}I$	$-\frac{3}{4}I^3$
NAND	-0.189	$-\frac{1}{2}I$	$-\frac{1}{2}I$	$-\frac{1}{2}I$	$-\frac{3}{4}I$	$-\frac{3}{16}I^3$
NOR	-0.189	$\frac{1}{2}I$	0	I	$\frac{3}{4}I$	$\frac{3}{4}I^3$

Table 3.3: While the interactions leave some gates indistinguishable, the J-outeractions of the input nodes are unique to each gate. As before: $I = 4 \log \frac{p}{\epsilon}$.

the output of an AND gate. Since the interactions are defined in a context of 0s, they order synergy with respect to observing 0s.

3.3.3 Interactions reflect dynamics beyond causal, correlation, or information quantities

3.3.3.a Interactions distinguish causal dynamics among triplets

I next investigated more complex causal dynamics than simple logic gates, to see how different association metrics reflect the underlying causal dynamics. I simulated different causal dynamics on the various 3-node causal DAGs shown in Figure 3.8. On a given DAG $\mathcal{G}$, denote the set of nodes without parents, the orphan nodes, by S_0. Each orphan node in S_0 got assigned a random value drawn from a Bernoulli distribution, i.e. $P(X = 1) = p$ and $P(X = 0) = 1 - p$. Denote the set of children of orphan nodes as S_1. Each node in S_1 then got assigned either the product of its parent nodes (for *multiplicative* dynamics), or the mean of its parent nodes (for *additive* dynamics), plus some zero-mean Gaussian noise with variance σ^2. All nodes were then rounded to a 0 or 1. A set S_2 was then defined as the set of all children of nodes in S_1, and these got assigned a value using the same dynamics as before. As long as the causal structure is acyclic, this algorithm terminates on a set of nodes S_i that has no children. For example, the chain graph $A \to B \to C$ has $S_0 = \{A\}$, $S_1 = \{B\}$, $S_2 = \{C\}$, and $S_3 = \emptyset$, at which point the updating terminates and the final sample is composed of the joint state of all nodes.

Figure 3.8 shows the different association metrics inferred on 100,000 states from each of the causal dynamics[6]. The six different dynamics corresponded to 4 different DAGs, 2 different correlation matrices, 4 different partial correlation matrices, and 2 different mutual information structures, which meant that each of these descriptions was degenerate in some of the dynamics. Partial correlations came close to disentangling direct from indirect effects, but failed to distinguish additive from multiplicative dynamics. I focused on the sign of the association and its significance, since the precise value depends on the noise level σ^2, but the precise values are listed in Appendix 3.C. The rightmost

[6] Multiplicative and additive dynamics are the same for colliderless graphs, like chains and forks.

column of Figure 3.8 shows that only the MFIs assigned a unique association structure to each of the dynamics, distinguished between direct and indirect effects, and revealed multiplicative dynamics as a 3-point interaction. Note that both the partial correlation and the MFIs assigned a negative association to the parent nodes in a collider structure. This reflects that two nodes become dependent when conditioned on a common effect (*cf.* Berkson's paradox), a spurious effect already found in partial correlations of metabolomic data in [140].

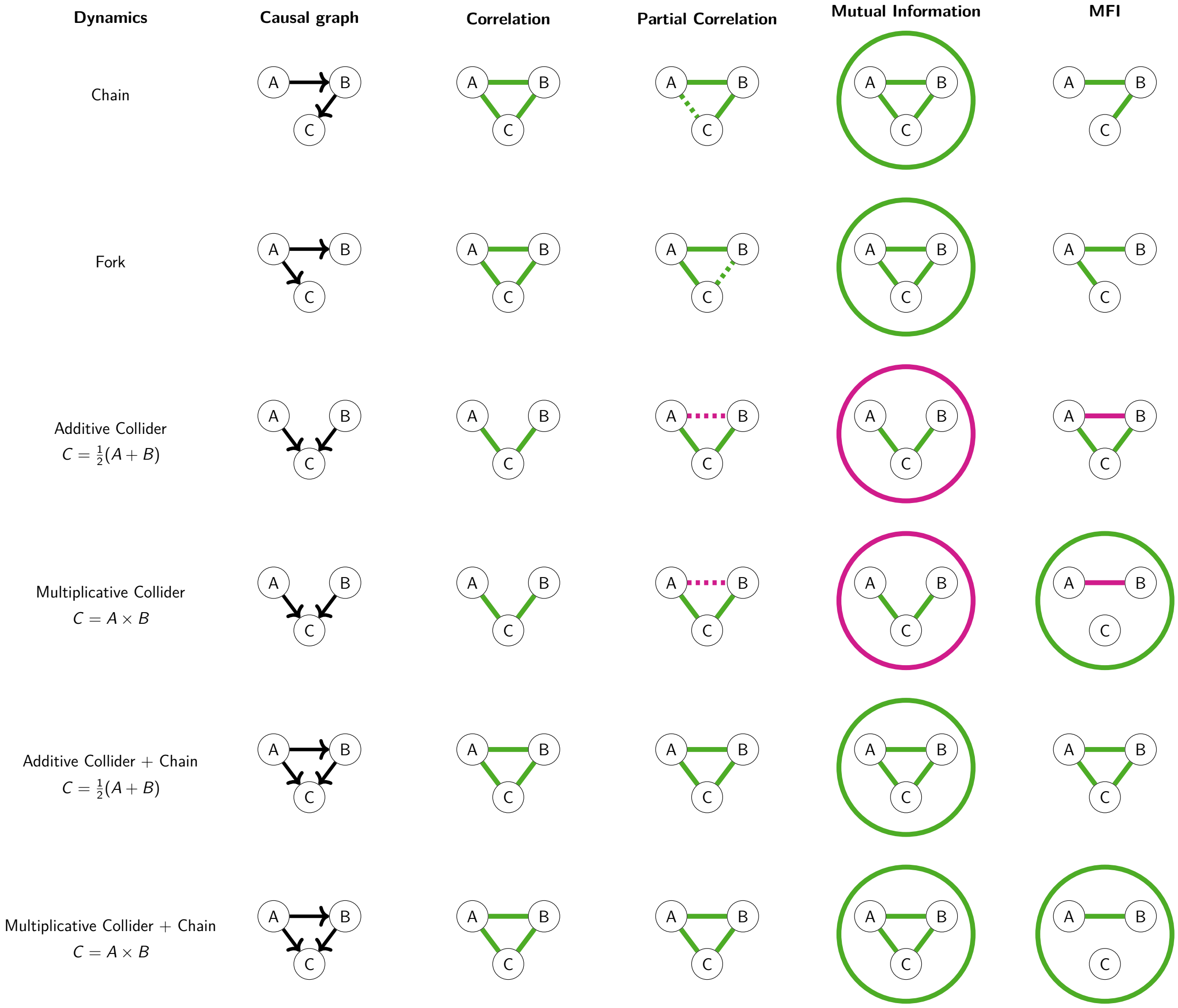


Figure 3.8: Different causal dynamics lead to different association metrics, and only MFIs can distinguish all 6 scenarios and reveal the combinatorial effect of a multiplicative interaction. Green edges denote positive values, red edges denote negative values, circles denote a 3-point quantity, and dashed lines show edges that show marginal significance that depends on σ^2. Correlations and mutual information cannot distinguish between most dynamics, and while partial correlation can, for certain noise levels, identify the correct pairwise relationships, it falls short of distinguishing additive from multiplicative dynamics. See Appendix 3.C for the simulation parameters and precise values.

3.3.3.b Simulations on larger DAGs

To see if the results from the previous section translate to more complex causal structures, I looked at larger DAGs that contain multiple triplet motifs. For each of the graphs in

Table 3.10, I simulated 10k graph configurations using the same procedure as in the previous section (setting $p = \sigma = 0.7$). The corresponding interactions are also listed in Table 3.10. The Tree and Collider Tree graphs were relatively straightforward, and reproduced the patterns of the isolated chains and colliders. More interesting was the confounded chain. It is a chain with a confounding node that leads to two colliders. Treating the chains and colliders in this DAG as isolated triplets leads to conflicting interactions, so the intuition about individual triplet motifs does not apply. The triplet $0 \to 1 \to 2$ is a chain, but had a 3-point coupling, and a pairwise coupling only between nodes 0 and 1. The same was true for the chain $0 \to 2 \to 3$: It had a 3-point interaction and only one 2-point interaction. Note that the chain $1 \to 2 \to 3$ had no interactions at all. Note also that while both colliders had a 3-point and a parent coupling, the parent coupling here was positive, while for isolated colliders it is negative. Simulations of additive dynamics on these DAGs yielded pairwise interactions only and did not conflict with intuition on the triplet motifs.

These results lead to the conclusion that the triplet motifs can offer intuition, but larger DAGs are more complex and lead to patterns in the interactions that cannot be reduced to those of colliders and chains by themselves.

3.3.3.c Interactions distinguish the dy- and triadic distributions

That the interactions have such resolving power over distributions of binary variables is perhaps not so surprising in light of the universality of RBMs with respect to this class of distributions [169, 84]. More surprisingly, their resolving power extends to the case of categorical variables. In [124], the authors introduce two distributions, the dy- and triadic distributions, that are indistinguishable by all Shannon-like information measures (in fact, they are indistinguishable by at least the following: Shannon-, Renyi(2)-, residual-, and Tsallis entropy, co-information, total correlation, CAEKL mutual information, interaction information, Wyner-, exact-, functional-, and MSS common information, perplexity, disequilibrium, and the LMRP- and TSE complexity). In this section, I will show that the MFIs perfectly distinguish these two distributions, and reveal the higher-order nature of the triadic distribution.

The two distributions are defined on 3 variables, each taking a value in a 4-letter alphabet $\{0, 1, 2, 3\}$. The joint probabilities are summarised in Table 3.11. To construct the distributions, each category is represented as a binary string— $(0, 1, 2, 3) \to (00, 01, 10, 11)$ — leading to new variables $\{X_0, X_1, Y_0, Y_1, Z_0, Z_1\}$. The dyadic distribution is constructed by linking these new variables with pairwise rules: $X_0 = Y_1, Y_0 = Z_1, Z_0 = X_1$, while the triadic distribution is constructed with rules involving triplets: $X_0 + Y_0 + Z_0 = 0 \mod 2$, and $X_1 = Y_1 = Z_1$. The resulting binary strings are then reinterpreted as categorical variables to produce Table 3.11. The authors of [124] find that no Shannon-like measure can distinguish between the two distributions, and they argue that the partial information decomposition, which is different for the two distributions, is not a natural information measure since it has to single out one of the variables as an output. To calculate model-free categorical interactions between the variables, I set the probabilities of the states in Table 3.11 uniformly to $p = (1-(64-8)\epsilon)/8$, and of the other states to ϵ (resulting in a normalised uniform distribution over legal states). There are a total of $6^3 = 216$ interactions such that $x_1 > x_0, y_1 > y_0, z_1 > z_0$. Each of these can be written

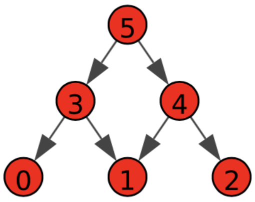


Table 3.4: Tree

	genes	Interaction	F
0	[5, 3]	0.287	0.000
1	[5, 0]	-0.001	0.449
2	[3, 0]	0.282	0.000
3	[3, 1]	-0.003	0.447
4	[3, 4]	-0.148	0.000
5	[0, 1]	0.019	0.042
6	[0, 2]	0.004	0.361
7	[5, 3, 4]	-0.001	0.497
8	[3, 0, 1]	0.008	0.342
9	[3, 4, 1]	0.320	0.000
10	[5, 3, 0]	0.009	0.316
11	[5, 3, 1]	-0.021	0.188

Table 3.5: Tree

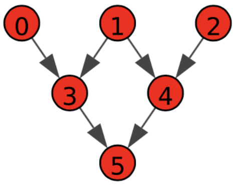


Table 3.6: Collider tree

	genes	Interaction	F
0	[0, 1]	-0.159	0.000
1	[0, 2]	0.032	0.001
2	[0, 3]	0.004	0.312
3	[0, 5]	0.017	0.060
4	[3, 4]	-0.112	0.000
5	[3, 5]	0.001	0.482
6	[0, 1, 2]	-0.027	0.028
7	[0, 1, 3]	0.305	0.000
8	[0, 3, 5]	-0.005	0.408
9	[0, 3, 4]	-0.002	0.453
10	[3, 4, 5]	0.261	0.000
11	[0, 2, 5]	-0.018	0.200

Table 3.7: Collider tree

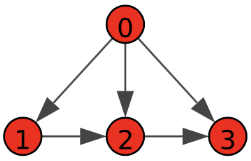


Table 3.8: Confounded Chain

	genes	Interaction	F
0	[0, 1]	0.145	0.000
1	[0, 2]	-0.124	0.000
2	[0, 3]	-0.006	0.266
3	[1, 2]	-0.002	0.433
4	[2, 3]	0.000	0.515
5	[0, 1, 2]	0.264	0.000
6	[0, 1, 3]	0.001	0.490
7	[0, 2, 3]	0.281	0.000
8	[1, 2, 3]	-0.010	0.344

Table 3.9: Confounded Chain

Table 3.10: Multiplicative DAGs and their interactions. F-values are bootstrapped.

as:

$$
\begin{aligned}
&I_{XYZ}(x_0 \to x_1; y_0 \to y_1; z_0 \to z_1) = \\
&\log \frac{p\Big(X = x_1, Y = y_1, Z = z_1 \mid \underline{X} = 0\Big)}{p\Big(X = x_0, Y = y_0, , Z = z_0 \mid \underline{X} = 0\Big)} \frac{p\Big(X = x_1, Y = y_0, Z = z_0 \mid \underline{X} = 0\Big)}{p\Big(X = x_0, Y = y_1, , Z = z_1 \mid \underline{X} = 0\Big)} \\
&\times \frac{p\Big(X = x_0, Y = y_1, Z = z_0 \mid \underline{X} = 0\Big)}{p\Big(X = x_1, Y = y_0, , Z = z_1 \mid \underline{X} = 0\Big)} \frac{p\Big(X = x_0, Y = y_0, Z = z_1 \mid \underline{X} = 0\Big)}{p\Big(X = x_1, Y = y_1, , Z = z_0 \mid \underline{X} = 0\Big)}
\end{aligned}
\tag{3.81}
$$

Of particular interest were two quantities: the interaction between the two most extreme states $I_{XYZ}(0 \to 3; 0 \to 3; 0 \to 3)$, and the symmetrised interaction $\bar{I}_{XYZ} = \sum_{x_0,x_1,y_0,y_1,z_0,z_1} I_{XYZ}(x_0 \to x_1; y_0 \to y_1; z_0 \to z_1)$, where the sum goes over all values such that $x_1 > x_0, y_1 > y_0, z_1 > z_0$, since all possible pairs necessarily sum to zero as $I_{XYZ}(x_0 \to x_1; y_0 \to y_1; z_0 \to z_1) = -I_{XYZ}(x_1 \to x_0; y_0 \to y_1; z_0 \to z_1)$.

For the dyadic distribution, I found:

$$
I^{\text{Dy}}_{XYZ}(0 \to 3; 0 \to 3; 0 \to 3) = \log \frac{p\epsilon^3}{p\epsilon^3} = 0 \tag{3.82}
$$

While for the triadic distribution:

$$
I^{\text{Tri}}_{XYZ}(0 \to 3; 0 \to 3; 0 \to 3) = \log \frac{\epsilon^4}{p\epsilon^3} = \log \frac{\epsilon}{p} \tag{3.83}
$$

So this particular 3-point interaction is zero for the dyadic, and negative for the triadic distribution. The sum over all 3-points is (see Appendix 3.B for details):

$$
\bar{I}^{\text{Dy}}_{XYZ} = \log 1 = 0 \tag{3.84}
$$

$$
\bar{I}^{\text{Tri}}_{XYZ} = 64 \log \frac{\epsilon}{p} \tag{3.85}
$$

That is, the symmetrised 3-point interaction is zero for the dyadic distribution, and strongly negative for the triadic distribution. These two distributions that are indistinguishable in terms of information structure are distinguishable by their model-free interactions, and these accurately reflect the higher-order nature of the triadic distribution.

3.4 Discussion

In this chapter, I defined and studied an estimator for model-free interactions (MFIs) that is used in the next chapters to investigate higher-order dependencies in gene expression data. I found that MFIs can be seen as a pointwise information quantity, similar

Dyadic

X	Y	Z	P
0	0	0	1 / 8
0	2	1	1 / 8
1	0	2	1 / 8
1	2	3	1 / 8
2	1	0	1 / 8
2	3	1	1 / 8
3	1	2	1 / 8
3	3	3	1 / 8

Triadic

X	Y	Z	P
0	0	0	1 / 8
1	1	1	1 / 8
0	2	2	1 / 8
1	3	3	1 / 8
2	0	2	1 / 8
3	1	3	1 / 8
2	2	0	1 / 8
3	3	1	1 / 8

Table 3.11: The joint probability of the dy- and triadic distributions (from [124]). All other states have probability zero.

to pointwise mutual information. On the lattice of subsets of variables, mutual information, pointwise mutual information, and MFIs appeared as the Möbius inversion of the expected, marginal, and joint surprisal, respectively. Furthermore, the dual quantities that arose had natural interpretations: dual mutual information corresponds to conditional, or differential, mutual information, while dual interactions are interactions in a context of 1s instead of 0s. Why Möbius inversions capture the structure associated with higher-order structure is not obvious. Möbius functions, with their recursive definitions, seem to capture the mereological[7] structure of their underlying poset, and as such play an important role in combinatorics. In fact, Möbius functions on Boolean algebras generalise the inclusion-exclusion principle that describes the number of elements in a union of sets. However, the role of Möbius inversions in higher-order information theory extends beyond Boolean algebras. In [101], the authors show that the lattice of antichains of subsets—not a Boolean algebra—captures the meronomy of higher-order information redundancy. As first noted by [300], the Möbius inversion over this lattice recapitulates the partial information decomposition that separates the synergistic, unique, and redundant information among variables.

As MFIs are defined in a completely model-free manner, they do not have a single intuitive or operational interpretation. I calculated the interactions up to third-order in different theoretical or simulated distributions to see how the MFIs reflected the underlying dynamical rules. On logic gates, third-order interactions corresponded to what is usually called synergy: the extent to which the output is unknown under incomplete knowledge of the input. Logical synergy can be captured by other higher-order quantities as well, like total correlation [222] and mutual information (Table 3.3). However, only the MFIs—in particular a derived quantity based on their dual—could perfectly distinguish each of the six possible logic gates.

A similar pattern was seen in noisy simulated dynamics of binary variables: only the MFIs could distinguish all simulated systems. Furthermore, the synergy present in multiplicative dynamics—approximately an AND-gate—was reflected by the presence of a 3-point interaction. In addition, the famously indistinguishable dy- and triadic distributions were

[7] Mereology is the study of the relationship between parts and wholes.

perfectly distinguished by their third-order MFIs which were only present in the triadic distribution. The dy- and triadic distributions are also distinguishable by their partial information decomposition, though only under the assumption that one of the variables is an output [124] while the MFIs are agnostic with respect to the causal direction.

In conclusion, in this chapter I have shown that the MFIs can be estimated by conditioning only on the Markov blanket, and can be assigned a confidence level. I also investigated their interpretation, and found that MFIs are closely related to information theory, but are better at disentangling direct from indirect effects and distinguishing logic and distributions. Finally, I found that only the MFIs accurately reflected higher-order dependencies in the data-generating process.

3.A Proofs

Proof of Lemma 2. *Let X be a set of variables with joint distribution $p(X)$. Let $A \in X$ and $B \in X$ such that $A \neq B$. Denote the minimal Markov blanket of X by MB_X. Then $A \in MB_B \iff B \in MB_A$, and we say that A and B are Markov-connected.*

Proof. Let $Y = X \setminus \{A, B\}$. Then

$$A \notin \mathrm{MB}_B \implies p(B \mid A, Y) = p(B \mid Y) \tag{3.86}$$

Consider

$$\begin{aligned} p(A \mid B, Y) &= \frac{p(A, B \mid Y)}{p(B \mid Y)} && (3.87)\\ &= \frac{p(B \mid A, Y) p(A, \mid Y)}{p(B \mid Y)} && (3.88)\\ &= p(A \mid Y) && (3.89) \end{aligned}$$

which means that $B \notin \mathrm{MB}_A$. Since $A \notin \mathrm{MB}_B \iff B \notin \mathrm{MB}_A$ holds, its negation also holds, which completes the proof. □

Proof of Proposition 1 *A model-free n-point interaction $I_{1\dots n}$ can only be nonzero when all variables $S = \{X_1, \dots, X_n\}$ are mutually Markov-connected.*

Proof. Let X be a set of variables with joint distribution $p(X)$. Let $S = \{X_1, \dots, X_n\}$, and $\underline{X} = X \setminus S$. Consider the definition of an n-point interaction among S:

$$\begin{aligned} I_{1\dots n} &= \prod_{i=1}^{n} \frac{\partial}{\partial X_i} \log p(X_1, \dots, X_n \mid \underline{X}) && (3.90)\\ &= \left(\prod_{i=1}^{n-1} \frac{\partial}{\partial X_i}\right) \frac{\partial}{\partial X_n} \log p(X_1, \dots, X_n \mid \underline{X}) && (3.91)\\ &= \left(\prod_{i=1}^{n-1} \frac{\partial}{\partial X_i}\right) \log \frac{p(X_n = 1 \mid X_1, \dots, X_{n-1}, \underline{X})}{p(X_n = 0 \mid X_1, \dots, X_{n-1}, \underline{X})} && (3.92)\\ &= \left(\prod_{i=1}^{n-1} \frac{\partial}{\partial X_i}\right) \log \frac{p(X_n = 1 \mid S \setminus X_n, \underline{X})}{p(X_n = 0 \mid S \setminus X_n, \underline{X})} && (3.93) \end{aligned}$$

Now, if $\exists X_j \in S \setminus X_n$ such that $X_j \notin \mathrm{MB}_{X_n}$, then then conditioning on X_j is not necessary as $p(X_n \mid S \setminus X_n) = p(X_n \mid S \setminus \{X_n, X_j\})$, so that

$$= \left(\prod_{i=1}^{n-1} \frac{\partial}{\partial X_i}\right) \log \frac{p(X_n = 1 \mid S \setminus \{X_j, X_n\}, \underline{X})}{p(X_n = 0 \mid S \setminus \{X_j, X_n\}, \underline{X})} \tag{3.94}$$

$$= \left(\prod_{\substack{i=1\\ i\neq j}}^{n-1} \frac{\partial}{\partial X_i}\right) \left(\frac{\partial}{\partial X_j} \log \frac{p(X_n = 1 \mid S \setminus \{X_j, X_n\}, \underline{X})}{p(X_n = 0 \mid S \setminus \{X_j, X_n\}, \underline{X})}\right) \tag{3.95}$$

$$= 0 \tag{3.96}$$

since the probabilities no longer involve X_j. Since X_j was chosen without loss of generality, this must hold for all variables in S, which means that if any variable in S is not in the Markov blanket of X_n, then the interaction I_S vanishes:

$$S \setminus X_n \not\subset \mathrm{MB}_{X_n} \implies I_S = 0 \tag{3.97}$$

Furthermore, the indexing of the variables was arbitrary, so this must hold for any reindexing, which means that

$$\forall X_i \in S : \quad S \setminus X_i \not\subset \mathrm{MB}_{X_i} \implies I_S = 0 \tag{3.98}$$

Which means that all variables in S must be Markov-connected for the interaction I_S to be nonzero. □

Proof of Proposition 1 *Let S be a set of random variables with probability distribution $p(S)$. Let X, Y, and Z be three disjoint subsets of S. Then omitting Y from the conditioning set results in a bias determined by, and linear in, the pointwise mutual information that $Y = 0$ gives about states of X:*

$$I_{X|YZ} - I_{X|Z} = \left(\prod_{i=1}^{|X|} \frac{\partial}{\partial x_i}\right) \mathrm{pmi}(X = x, Y = 0 \mid Z = 0) \tag{3.99}$$

Proof. The pointwise mutual information (pmi) is defined as

$$\mathrm{pmi}(X = x, Y = y) = \log \frac{p(X = x, Y = y)}{p(X = x)p(Y = y)} \tag{3.100}$$

Note that

$$p(X = x_1 \mid Y = y, Z = z) = \frac{p(X = x_1, Y = y \mid Z = z)}{p(Y = y \mid Z = z)} \tag{3.101}$$

So that

$$p(X = x_1 \mid Y = y, Z = z) = e^{\mathrm{pmi}(X = x_1, Y = y \mid Z = z)} p(X = x_1 \mid Z = z) \tag{3.102}$$

That is, not conditioning on $Y = y$ results in an error in the estimate of $p(X = x_1 \mid Y = y, Z = z)$ that is exponential in the Z-conditional pmi of X and Y. However, consider the interaction among X:

$$I_X = I_{X|YZ} = \left(\prod_{i=1}^{|X|} \frac{\partial}{\partial x_i}\right) \log p(X = x \mid Y = 0, Z = 0) \tag{3.103}$$

$$= \left(\prod_{i=1}^{|X|} \frac{\partial}{\partial x_i}\right) (\log p(X = x \mid Z = 0) + \text{pmi}(X = x, Y = 0 \mid Z = 0)) \tag{3.104}$$

$$= I_{X|Z} + \left(\prod_{i=1}^{|X|} \frac{\partial}{\partial x_i}\right) \text{pmi}(X = x, Y = 0 \mid Z = 0) \tag{3.105}$$

That is, the error in the interaction as a result of not conditioning on the right variables is linear in the difference between the pmi's of different states. □

3.B Python code to calculate categorical dy- and triadic interactions

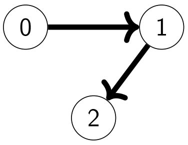
Code 3.1: calculate dy- and triadic MFIs

```
dyadicStates = [[’a’, ’a’, ’a’], [’a’, ’c’, ’b’], [’b’, ’a’, ’c’], [’b’, ’c’, ’d’],
[’c’, ’b’, ’a’], [’c’, ’d’, ’b’], [’d’, ’b’, ’c’], [’d’, ’d’, ’d’]]

triadicStates = [[’a’, ’a’, ’a’], [’a’, ’c’, ’c’], [’b’, ’b’, ’b’], [’b’, ’d’, ’d’],
[’c’, ’a’, ’c’], [’c’, ’c’, ’a’], [’d’, ’b’, ’d’], [’d’, ’d’, ’b’]]

stateDict = {0: ’a’, 1: ’b’, 2:’c’, 3: ’d’}

def catIntSymb(x0, x1, y0, y1, z0, z1, states):
    prob = lambda x, y, z: ’p’ if [x, y, z] in states else ’e’

    num = prob(x1, y1, z1) + prob(x1, y0, z0) + prob(x0, y1, z0) + prob(x0, y0, z1)
    denom = prob(x1, y1, z0) + prob(x1, y0, z1) + prob(x0, y1, z1) + prob(x0, y0, z0)
    return (num, denom)

numDy = ’’
denomDy = ’’
numTri = ’’
denomTri = ’’

for x0 in range(4):
    for x1 in range(x0+1, 4):
        for y0 in range(4):
            for y1 in range(y0+1, 4):
                for z0 in range(4):
                    for z1 in range(z0+1, 4):

                        nDy, dDy = catIntSymb(*[stateDict[x] for x in [x0, x1, y0, y1, z0, z1]], dyadicStates)
                        numDy += nDy
                        denomDy += dDy

                        nTri, dTri = catIntSymb(*[stateDict[x] for x in [x0, x1, y0, y1, z0, z1]], triadicStates)
                        numTri += nTri
                        denomTri += dTri

```

```

print(f'Dyadic interaction: log (p^{numDy.count("p") - denomDy.count("p")} e^{numDy.
    count("e") - denomDy.count("e")})')
print(f'Triadic interaction: log (p^{numTri.count("p") - denomTri.count("p")} e^{numTri.
    count("e") - denomTri.count("e")})')

// Output:

>> Dyadic interaction: log (p^0 e^0)
>> Triadic interaction: log (p^-64 e^64)
```

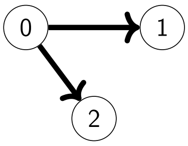
3.C Numerics of causal structures

Tables 3.12 to 3.17 list the precise values that led to Figure 3.8. From each graph, I generated 100k samples using $p = 0.5$ and $\sigma = 0.4$. To quantify the significance value of the interactions, I generated 1,000 bootstrap resamples of the data, and calculated F: the fraction of resampled interactions that have a different sign from the original interaction. The smaller F is, the more significant the interaction is.



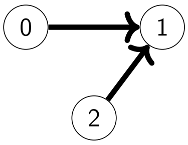
	Genes	Interaction	F	ρ_p	p-val	Partial ρ_p	p-val	MI
0	[0, 1]	4.281	0.000	0.790	0.0	0.635	0.000e+00	0.515
1	[0, 2]	0.056	0.117	0.622	0.0	0.031	2.261e-23	0.301
2	[1, 2]	4.249	0.000	0.786	0.0	0.628	0.000e+00	0.510
3	[0, 1, 2]	-0.052	0.217	NaN	NaN	NaN	NaN	0.300

Table 3.12: Chain



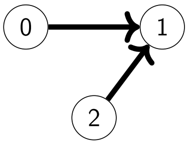
	Genes	Interaction	F	ρ_p	p-val	Partial ρ_p	p-val	MI
0	[0, 1]	4.268	0.000	0.789	0.0	0.634	0.000e+00	0.514
1	[0, 2]	4.257	0.000	0.788	0.0	0.632	0.000e+00	0.512
2	[1, 2]	-0.014	0.376	0.622	0.0	0.028	6.518e-19	0.300
3	[0, 1, 2]	0.020	0.376	NaN	NaN	NaN	NaN	0.300

Table 3.13: Fork



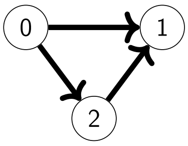
	Genes	Interaction	F	ρ_p	p-val	Partial ρ_p	p-val	MI
0	[0, 1]	2.144	0.000	0.395	0.000	0.505	0.000e+00	1.154e-01
1	[0, 2]	-0.989	0.000	-0.002	0.593	-0.070	5.172e-109	2.059e-06
2	[1, 2]	2.144	0.000	0.395	0.000	0.505	0.000e+00	1.154e-01
3	[0, 1, 2]	0.003	0.438	NaN	NaN	NaN	NaN	-2.678e-02

Table 3.14: Additive collider



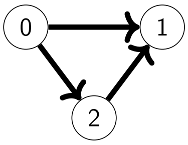
	Genes	Interaction	F	ρ_p	p-val	Partial ρ_p	p-val	MI
0	[0, 1]	0.032	0.140	0.427	0.000	0.478	0.000e+00	1.403e-01
1	[0, 2]	-2.156	0.000	-0.005	0.145	-0.087	1.463e-166	1.529e-05
2	[1, 2]	0.036	0.109	0.429	0.000	0.480	0.000e+00	1.415e-01
3	[0, 1, 2]	4.237	0.000	NaN	NaN	NaN	NaN	-1.150e-01

Table 3.15: Multiplicative collider



	Genes	Interaction	F	ρ_p	p-val	Partial ρ_p	p-val	MI
0	[0, 1]	2.103	0.000	0.705	0.0	0.362	0.0	0.396
1	[0, 2]	3.288	0.000	0.790	0.0	0.599	0.0	0.515
2	[1, 2]	2.113	0.000	0.706	0.0	0.364	0.0	0.397
3	[0, 1, 2]	0.050	0.162	NaN	NaN	NaN	NaN	0.335

Table 3.16: Additive collider + chain



	Genes	Interaction	F	ρ_p	p-val	Partial ρ_p	p-val	MI
0	[0, 1]	-0.017	0.342	0.709	0.0	0.365	0.0	0.403
1	[0, 2]	2.094	0.000	0.786	0.0	0.596	0.0	0.510
2	[1, 2]	-0.057	0.092	0.707	0.0	0.361	0.0	0.401
3	[0, 1, 2]	4.359	0.000	NaN	NaN	NaN	NaN	0.293

Table 3.17: Multiplicative collider + chain

Chapter 4

Higher-order mechanism and regulation

(causal thinking never yields accurate descriptions of metabolic processes—limitations of existing language)

William S. Burroughs [42]

4.1 Introduction

In this chapter, I investigated to what extent the higher-order dependencies in observational gene expression data reflect underlying biological mechanism. To do so, I calculated MFIs on scRNA-seq data sets from mouse brains at different stages of development. This allowed me to see how different biology reveals itself in different statistical dependencies. As this is the first time that MFIs have been calculated on biological data, I first investigated how robust the estimates were with respect to changes in the underlying expression data, before validating the MFIs against known biology. Validation of model-free quantities is a subtle matter, and there is not always a clear separation between validation and discovery, as I will emphasise in the following sections.

4.1.1 Validation of model-free interactions

Parameters in a model generally have a clear interpretation since models have to be theoretically justified and usually try to be parsimonious to preserve statistical power. This interpretation makes validation relatively straightforward: if regression coefficients of a particular model disagree with known interactions, for example, then the model that produced them can be refuted. For model-free quantities, interpretation—and by extension validation—tends to be more difficult. For example, while calculating correlations is trivial, their interpretation depends on the context in which they were calculated and can be notoriously misleading [191, 92, 197, 291]. If a measured correlation cannot be explained, there is nothing to refute but the data. This makes it hard to see how a model-free quantity could lead to a testable hypothesis, and places the emphasis instead on interpreting the quantity being estimated. This distinction is crucial, and will guide the validation of the MFIs in the rest of this thesis.

The model-free interactions are defined with respect to the joint probability of a transcriptional state, so they reflect how different expression patterns have different probabilities. Therefore, a non-zero MFI can be explained at three different levels, each corresponding to a different cause of the relative likelihoods of the expression patterns. At the most basic level, biochemical interaction networks can preferentially produce certain RNA molecules over others, leading to the observed differences in RNA abundance. For example, a transcription factor A binding to the promoter of a gene B can increase the production of B-transcripts. In the absence of the effect of other genes and autoregulation, and under the common assumption that RNA and protein abundance are correlated[1], more A-transcripts should lead to more A-proteins, which leads to more B-transcripts. This would manifest itself in a positive 2-point interaction $I_{AB} > 0$ at the level of RNA. In this case, the interaction would be directly interpretable in terms of a biological mechanism, namely a transcription factor binding to its target promoter. At a higher level of abstraction, the different expression patterns in the data could be caused by the presence of distinct cell states. For example, a certain stressor in the environment of some of the cells might trigger the rapid response of n different genes, all of which increase their rate of transcription within minutes [16]. Even if the response genes do not interact biochemically with each other, this conditional expression pattern would—*by definition*—result in a non-zero n-point MFI. These different cell states would still be

[1] This is, however, not always the case, as discussed in Section 1.4 and [26, 285]

the result of biochemical mechanisms [66, 16], but the relative likelihood of expression patterns would reflect the relative proportions of the different states, rather than the biochemical mechanisms directly. Finally, the different expression patterns could correspond to cells of different types. When inferring interactions on a data set that contains both astrocytes and neurons, for example, the relative likelihood of expression patterns could be dominated by the difference in expression of marker genes for neurons and astrocytes. Each of these three possibilities lends a different interpretation to the MFIs, and should be validated differently. Note that they are not disjoint explanations—cell state transitions are driven by biochemical mechanisms, and there is no clear distinction between cell state and cell type. Most likely, the interactions will reflect a combination of these three different structures. Still, I will explore these modes of explanation separately, and start with exploring the mechanistic content of the estimated interactions, which is the focus of this chapter.

4.1.2 Interactions as biological mechanism

The hypothesis that MFIs reflect biological mechanisms would be supported by finding that the interactions agree with known biological pathways and preferentially occur between genes whose proteins have previously been annotated to physically interact. To investigate this possibility, I compared the network of MFIs with the association networks from various established *gold standard* biological sources, as collected in the Pathway Commons database [219]. These Pathway Commons associations are separated into different categories of association—regulatory, protein binding, phosphorylation, *etc.*—so seeing which category the MFIs most closely resemble can help develop a mechanistic interpretation of the MFIs. Still, validation is not straightforward. These databases of associations are notoriously incomplete [118, 123, 17], so they are mainly useful to identify true positives. Moreover, the databases tend to be organism-wide, or even integrate knowledge across different organisms and throughout developmental stages, whereas I inferred interactions on only a few cell types from the mouse brain, at only one developmental time point per estimation. The interactions in the database that only occur in other cell types, tissues or organisms should therefore not result in an MFI, but would incorrectly show up as false negatives in a naive validation. In addition, databases of interactions commonly focus only on protein-protein interactions, while the MFIs reflect dependencies at the transcriptional level. Furthermore, the MFIs are defined (for the purpose of this thesis) in a context of unexpressed genes, whereas gold standard databases aim to capture the *in vivo* interactions that occur in a diverse and dynamic transcriptomic context, which might lead to different conclusions. Finally, the construction of gold standard networks is also known to be biased in different ways. Since genes that show significant coexpression patterns are more likely to be the subject of a follow-up study, many databases are biased towards sets of genes that correlate strongly in at least some tissues [94], and genes that are already known to play a role in important pathways or disease tend to be overstudied and overrepresented in gold standard networks [228].

An orthogonal measure of mechanistic association is in terms of gene ontology. A gene ontology is a DAG where the nodes are biological annotations, whose specificity increases along the orientation of the edges. In a method first described in [112], a gene gets assigned to nodes in the DAG, and a pair of genes can be assigned their last common ancestor's node. The more specific their last common ancestor's annotation is, the more *semantically similar* the two genes are. Depending on what kind of biology the nodes in the ontology graph describe, one can assign different kinds of semantics to the similarity. For example, in the ontology graph where nodes are subcellular locations and genes get assigned their products' primary location, semantic similarity quantifies the colocalisation of gene products. The concept of semantic similarity and different gene ontologies will be described in more detail in Section 4.2.2. Concordance with mechanistic ontologies can serve as further evidence that the MFIs reflect mechanisms.

4.1.3 Aim and outline of this chapter

This chapter aims to describe the relationship between biological mechanism and higher-order dependencies in gene expression data, as revealed through model-free interactions. To establish a biological 'ground truth' **Section 4.2.1** introduces the Pathway Commons database as a 'gold standard' network, and **Section 4.2.2** discusses the various ways ontologies can be used to quantify biological relationships. Before the MFIs can be calculated, all Markov blankets have to be estimated, so their definition is recapitulated in **Section 4.2.3**. Their estimation requires the quasi-causal graph of conditional dependencies, and two different causal discovery algorithms are introduced: the Peter-Clark algorithm (**Section 4.2.4**) and an iterative MCMC procedure (**Section 4.2.5**). **Section 4.2.6** introduces and describes the different data sets under consideration: mouse neurons and astrocytes at two developmental stages.

Before validating the MFIs against biology, I first investigated how much data is needed to estimate the MFIs in **Section 4.3.1**, and how robust the estimation procedure is with respect to changing the cells (**Section 4.3.2**) and the genes (**Section 4.3.3**) in the data set. I then quantified the biological content of the MFIs by first comparing functional enrichment of the different orders of interactions in **Section 4.3.4**. **Section 4.3.5** shows that the MFIs disentangled direct from indirect effects. **Section 4.3.6** compares the semantic similarity of correlated and interacting genes, and **Section 4.3.7** explores how higher-order interactions relate to combinatorial transcription factor binding. **Section 4.3.8** validates individual MFIs and coexpression networks against the Pathway Commons database, **Section 4.3.9** associates higher-order interactions to genetic logic gates, and **Section 4.3.10** shows how the MFI networks were modular with respect to protein function. Finally, **Section 4.4** reflects on the results from this chapter, and motivates the transition to state inference in **Chapter 5**.

4.2 Methods

4.2.1 Validation against 'gold standard' networks

To validate the MFIs against known biological mechanisms, I compared the network of MFIs with the Pathway Commons database, which aggregates human pathways and mechanistic relationships across multiple sources, including the Reactome and Kegg databases [219]. However, these databases are incomplete, not tissue-specific, and conflate different kinds of molecular mechanisms. Consequently, these databases mostly identify true positives, as false positives might be due to the incompleteness of the databases, and false negatives due to tissue-specificity, imperfect orthology mapping, or because the ground truth molecular association does not affect RNA concentrations. Consequently, I only quantify the performance by the *precision* of the MFIs relative to Pathway Commons, and not the *recall*. Pathway Commons categorises the associations by mechanism, and I focused on the following categories:

1. A controls the expression of B.
2. A controls a state change of B (*e.g.* through a post-translational modification of the protein B).
3. A is in a complex with B.
4. A interacts with B (*i.e.* both participate in a molecular interaction according to a particular kind of digital bioinformatics object known as PSI-MI [132]).
5. Miscellaneous: all other annotations.

More details on these categories can be found at `https://www.pathwaycommons.org/pc2/formats`. I downloaded version 12 of the Pathway Commons database from `pathwaycommons.org/archives/PC2/v12` (last modified on September 18, 2019), and used `gProfiler`'s orthology API (`g:Orth`) to map human genes to mouse genes. Since MFIs are symmetric, I symmetrised the Pathway Commons database, making each association bidirectional. A true positive is then defined as an unordered pair of genes (A, B) that appears as a pair in the symmetrised Pathway Commons database and also appears together in an MFI.

4.2.2 Validation against gene ontologies

4.2.2.a Ontological enrichment

Annotating genes and their products with their biological function is not straightforward, as one gene can have many distinct roles, and multiple genes can have similar roles. Furthermore, descriptions of biological function are not independent but instead form a hierarchy. For example, the genes that aid with cell cycle progression are a superset of those that regulate DNA replication, which are a superset of those that play a role in ‘mitotic recombination-dependent replication fork processing’ (in mice these are the genes *Brca2* and *Rad51*). However, the full hierarchy is not necessarily tree-like, as a particular function can be part of multiple processes higher up in the hierarchy. The *Gene Ontology* consortium (GO, [15, 3]) aims to standardise these descriptive terms and their hierarchy. As of September 2022, there are 43,558 terms included in the GO database, with 7,483,496 annotations across 1,480,259 gene products, across 5,213 species. These terms and annotations are distributed across three disjoint graphs, each describing a particular domain of annotations. The biological process graph (GO:BP) contains 28,140 terms and describes “the larger processes, or ‘biological programs’ accomplished by multiple molecular activities”, *i.e.* more abstract processes that do not necessarily correspond to mechanistic pathways. The Molecular Function (GO:MF) graph contains 11,238 terms and describes “activities that occur at the molecular level, such as ‘catalysis’ or ‘transport’”. Finally, the Cellular Component (GO:CC) graph contains 4,180 terms and describes “The locations relative to cellular structures in which a gene product performs a function", *i.e.* the terms correspond to locations in the cell, not to gene functions.

A set of genes thus has a corresponding set of associated ontology terms. Given a set S of genes, terms that appear more often than expected by chance are called *enriched*. A term t—seen as the set of all associated genes—has a certain size: the total number of genes annotated to it, and under the null hypothesis that S is a uniform and independent sample from the set of all genes G, the size of the intersection $S \cap t$ follows a hypergeometric distribution. That is, the intersection size is distributed as

$$p(|S \cap t| = k) = \frac{\binom{|t|}{k}\binom{|G|-|t|}{|S|-k}}{\binom{|G|}{|S|}} \tag{4.1}$$

Observed intersection sizes can be used to assign a significance to the enrichments, and an m-fold enrichment corresponds to the situations where

$$\frac{|S \cap t|}{\mathbb{E}_p[k]} = m \tag{4.2}$$

where $\mathbb{E}_p[k]$ is the expected value of k under the hypergeometric distribution in Equation (4.1). Using this, sets of genes can be understood through the terms they are enriched in. To quantify the enrichment of a set of genes, a background set of genes has to be defined, typically chosen to be the set of genes that *could* have ended up in the set of interest, although note that this definition can be ambiguous. Since the genes are chosen in a data-driven way from the set of all genes, one could argue that the background set should include all genes for which the gene expression was measured. However, I chose the more conservative background of only those genes that remained in the final MFI estimation step (typically a few hundred to a thousand genes in total).

Testing all gene sets for enrichment in all possible terms involves many tests, so the statistical significance threshold should be accordingly corrected. However, the validity of both Bonferroni correction and Benjamini-Hochberg false discovery rate analysis depends on the dependency among the hypothesis tests, so correcting p-values for multiple testing can be misleading when the terms have a hierarchical structure like in ontology graphs and pathways. To remedy this, I queried enrichment with the `gProfiler` API which applies a pre-computed `g:SCS` multiple hypothesis correction to each p-value that takes the hierarchical structure into account. `gProfiler` can do this for more than just gene ontologies, so I used `g:Profiler`'s API [211] to query enrichment in terms from the following databases: Gene Ontology (molecular function, biological process, and cellular component), KEGG pathways, Reactome pathways, WikiPathways, Transfac transcription factor binding site predictions, MirTarBase miRNA targets, CORUM protein complexes, and Human Phenotype ontology.

4.2.2.b Semantic similarity

Pairs of genes can also be compared in terms of their gene ontology. If two genes are somehow 'close' on the ontology graph, they might be mechanistically related as well. A concept known as semantic similarity makes intuition precise. Two genes can be called similar if their annotations share a specific ancestor on an ontology graph. Different graphs and definitions of specificity lead to different notions of similarity. One of the oldest and most common measures of the specificity of a term t—referred to as the Resnik method [215]—is the surprisal of that annotation, relative to all annotations in the ontology: $S_t = -\log \frac{|\tau(t)|}{N}$, where $\tau(t)$ is the set that includes the term t and all its descendants, and N is the total number of terms in the ontology. While not necessarily tree-like, the ontology graphs do form a well-defined hierarchy, so any set of terms will have a 'last common ancestor' term, namely the most informative common ancestor term (MICA, the common ancestor with the largest surprisal). The Resnik method defines the semantic similarity $\mathrm{sim}_{s,t}$ of two terms s and t as the surprisal of their MICA:

$$\mathrm{sim}_{s,t} = -\log \frac{|\tau(\mathrm{MICA}(s,t))|}{N} \tag{4.3}$$

However, a more modern method—Wang's method [289]—assigns a similarity score based on not just the MICA, but on all ancestors. To account for the relative distances to each of the terms, each ancestor gets weighted by a term that depends on the local topology. Throughout this thesis, I will use Wang's method to define the semantic similarity between ontology terms. Since genes often get assigned to multiple terms, the semantic similarity of two genes is an aggregate of the similarity of their corresponding terms—typically an average or the maximum. To aggregate the term similarities into a gene similarity, I use the so-called best-match average (BMA) method: Given two genes g_1 and g_2—with associated sets of terms S and T of size m and n, respectively—construct the $m \times n$ matrix M where $M_{ij} = \mathrm{sim}(S_i, T_j)$. The semantic similarity between g_1 and g_2 is then

$$\mathrm{sim}_{BMA}(g_1, g_2) = \frac{1}{m+n}\left(\sum_{i=1}^{n} \max_j M_{ij} + \sum_{j=1}^{m} \max_i M_{ij}\right) \tag{4.4}$$

This essentially matches terms from S and T so that the mutual similarity is maximised, and then sets the similarity of g_1 and g_2 to the mean across all these best matched pairs. In practice, I use the implementation of these similarity scores in the `GOSemSim` package [312].

4.2.3 Markov blankets and boundaries

The role of Markov blankets in the estimation of MFIs was introduced and discussed in Section 3.2.2. This section mainly concerns their practical estimation. If the joint distribution $p(X)$ is Markov compatible with a DAG $\mathcal{G}$, the minimal Markov blanket of a node X_i is composed of the parents, children, and co-parents, or spouses, of X_i. In other words: the Markov blanket of X_i is composed of all nodes reachable by a single step on $\mathcal{G}$, or a single step on $\mathcal{G}^{\mathrm{op}}$, or a step on $\mathcal{G}$ followed by a step on $\mathcal{G}^{\mathrm{op}}$, where $\mathcal{G}^{\mathrm{op}}$ is the graph $\mathcal{G}$ with all arrows reversed. Given the adjacency matrix $\mathcal{A}$ corresponding to $\mathcal{G}$, all Markov blankets are thus encoded in the matrix $\mathrm{MB} = \mathcal{A} + \mathcal{A}^T + \mathcal{A}^T\mathcal{A}$ (note that the symmetry implied by Lemma 2 is immediately obvious in this case: $\mathrm{MB} = \mathrm{MB}^T$). Note that if A is a Markov blanket for B, then $A \cup C$ is also a Markov blanket for B, for any C. That is, conditioning on more than the minimal Markov blanket might increase the variance of the estimator, but it does not introduce a bias. To estimate the matrix $\mathcal{A}$, I combined two causal discovery algorithms: the constraint-based Peter-Clark algorithm (covered in Section 4.2.4) and a score-based MCMC method (described in Section 4.2.5). It should be noted that inferring the causal graph from purely observational data is impossible in most cases. Therefore, the edges from the adjacency matrix should not be interpreted as true causal links. To emphasise this, I will refer to the associated graph as a *quasi*-causal graph throughout this thesis. This is not a limiting factor in the estimation of the interactions, as the graph is only used to decide on conditional dependencies. True causal discovery and regulatory inference generally require a mix of observational and interventional data, for example using large-scale perturb-seq experiments.

4.2.4 The Peter-Clark algorithm

Finding the graph of dependencies among n variables requires each pair to be tested so can be done with $\binom{n}{2} = n(n-1)/2$ dependency tests. In most cases, this $\mathcal{O}(n^2)$ scaling means it is a tractable problem. The scaling becomes much worse for conditional dependencies. There, for each pair, at worst 2^{n-2} conditional dependency tests have to be performed (one for each subset of the remaining nodes). This exponential scaling makes constructing conditional dependency graphs intractable in most cases. The PC algorithm mitigates this scaling by doing the dependency tests in a convenient order, iterating over levels of increasingly stringent conditioning, and eliminating possible dependencies at each iteration.

4.2.4.a Causal discovery with oracles, skeletons, and colliders

This section describes the structure of the Peter-Clark (PC) algorithm, assuming access to a *CD-oracle*: an imaginary source of perfect information on conditional dependence between any two variables. Replacing this oracle with practical hypothesis tests will be covered afterwards.

The algorithm starts in the most conservative way, by assuming that all variables are dependent on each other, *i.e.* with the fully connected (undirected) conditional dependency (CD) graph $\mathcal{G}$. This is most conservative as it leads to the largest Markov blankets possible. The algorithm then iterates and eliminates dependencies at increasingly high levels. At level 0, no conditioning is done, and the CD-oracle is consulted for each pair $\{X_i, X_j\}$. If $X_i \perp\!\!\!\perp X_j$, the edge between X_i and X_j is removed from $\mathcal{G}$. At level 1, only those pairs that are still connected in $\mathcal{G}$ are considered. For each pair $\{X_i, X_j\}$, a node Y that is still adjacent to X_i or X_j in $\mathcal{G}$ is chosen. If $X_i \perp\!\!\!\perp X_j \mid Y$, the edge between X_i and X_j is removed from $\mathcal{G}$. This is repeated for all neighbours of X_i and X_j. Then, at each higher level ℓ, instead of picking a single node Y, a set Z of ℓ neighbours of X_i or X_j is chosen. The edge between X_i and X_j is then removed if and only if $X_i \perp\!\!\!\perp X_j \mid Z$. This is done for all such subsets. This process terminates when a level $\ell^{\max}$ is reached such that no node has a degree $> \ell^{\max}$, at which point there are no more dependencies tests to do. At this point, $\mathcal{G}$ is an undirected graph called a *skeleton*. Note that this skeleton does not contain all information on conditional dependence. If the underlying DAG contains a *collider* structure $X \rightarrow Y \leftarrow Z$, then $X \perp\!\!\!\perp Z \mid Y$ would not hold, while if the underlying structure was a chain $X \rightarrow Y \rightarrow Z$ or a fork $X \leftarrow Y \rightarrow Z$, then $X \perp\!\!\!\perp Z \mid Y$ would hold. While the skeleton does not distinguish between these two scenarios, the conditional dependency structure does, and edges can be oriented accordingly by consulting the oracle for each potential collider triplet. Sometimes, with all colliders identified and oriented, more edges can be oriented by requiring that no other colliders exist other than the ones found in the previous step, and making sure that no oriented cycles appear. The PC algorithm terminates when all these have been oriented.

Even with a perfect dependency oracle, the DAG compatible with the joint distribution is thus only identifiable up to an equivalence class of completed partially oriented DAGs, or CPDAGs. Furthermore, the claim that $\mathcal{G}$ accurately captures the dependency structure of joint distribution $p(X)$ implicitly assumes that the pair $(p, \mathcal{G})$ has the following three properties:

1. Causal Markov property of the pair $(p(X), \mathcal{G})$: Each variable $X_i \in X$ is independent (*i.e.* the joint probability factorises) of all its non-descendants in $\mathcal{G}$, when conditioned on its parents in $\mathcal{G}$.
2. Faithfulness property of the pair $(p(X), \mathcal{G})$: The independencies implied by $\mathcal{G}$ are the only independencies that hold in $p(X)$.
3. Causal sufficiency property of $\mathcal{G}$: There are no unobserved variables with more than one descendant in X.

This is the PC algorithm as it was originally introduced. It suffers from a dependence on the order in which the tests are done since edges are immediately removed upon discovering an independency. This problem can be solved by removing edges only after doing all tests at a certain level, at the cost of doing more tests within each level. This order-independent version is the one used in this thesis and is generally referred to as the stable PC algorithm.

4.2.4.b Practical dependency tests

Unfortunately, CD-oracles do not exist. Identifying conditional independencies is instead done with hypothesis tests, and requires a significance threshold at which to reject the null hypothesis. For most commonly used tests, the null hypothesis is that there is no dependency. To be conservative, I choose a relatively large significance threshold $p \leq 0.05$ at which to reject the null hypothesis. Note that this is different from the normal use of statistical tests, in which a smaller α is more conservative. If the null hypothesis is not rejected, then it is concluded that $X_i \perp\!\!\!\perp X_j \mid Z$, and the edge between node i and j is removed from the CD-graph $\mathcal{G}$.

For categorical/binary variables, dependency tests can be carried out with only the contingency table of the samples. Traditionally a χ^2-statistic is constructed as

$$X^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i} \tag{4.5}$$

where the sum goes over all N joint configurations in the contingency table (*i.e.* $\{00, 01, 10, 11\}$ in the case of binary variables). Here O_i is the observed incidence of joint state i, and E_i is the expected incidence of joint state i, as predicted from the mean of each variable. This statistic follows a χ^2-distribution (with 2 degrees of freedom) under the null hypothesis of independence, and can thus be used to test for dependence. In practice, a G-test tends to perform better and is the one used by the R-package `pcalg` [55]. The statistic G is constructed as:

$$G = 2\sum_i O_i \cdot \ln\left(\frac{O_i}{E_i}\right) \tag{4.6}$$

which follows a χ^2-distribution with the same number of degrees of freedom more robustly in some cases. Note that this statistic is essentially the KL-divergence between the observed distribution and the distribution expected from the marginals.

4.2.4.c Ambiguity

Consider a potential collider triplet in the skeleton $A - B - C$. If it is a collider, then A and C are dependent, conditioned on B and all subsets of neighbours of A and C. In the baseline (*i.e.* non-stable) PC algorithm, this triplet is oriented as a collider if B was *not* part of the conditioning set that led to the removal of the edge between A and C. This introduces another source of order dependency. To deal with this, the authors of [209] propose an algorithm they call conservative PC (CPC) in which the triplet is only oriented as a collider structure when B is in *none* of the sets $Z \subseteq \text{adj}(A) \cup \text{adj}(B)$ such that $A \perp\!\!\!\perp C \mid Z$. If B is in all such separating sets, it is identified as a non-collider. Otherwise, the edge is left as *ambiguous*. This, however, is too conservative in most cases, and the authors of [55] propose what is called the majority rule, in which the triplet is oriented as a collider if B appears in fewer than half of all sets Z that make $A \perp\!\!\!\perp C \mid Z$, and identified as a non-collider otherwise[2]. Note that both CPC and the majority rule make the orientation procedure independent of the node ordering. Throughout this thesis, I orient edges with the majority rule using the stable PC algorithm as implemented in the R-package `ParallelPC` [146] that parallelises the `pcalg` package [55].

A final source of ambiguity is conflicting directions, *e.g.* when there is evidence for $A \to B \leftarrow C$ as well as $B \to C \leftarrow D$. The baseline PC algorithm just overwrites all edges, which reintroduces an order dependence here. Instead, I marked edges with conflicting directions as bidirectional, fully preserving order independence.

4.2.5 Optimising the causal graph with MCMC methods

The PC algorithm is an example of a constraint-based method for causal discovery—it evaluates DAGs by verifying the constraints implied by causal structures. In contrast, score-based methods assign each DAG a numerical value based on how well it fits the data. Score-based methods are computationally more expensive but tend to perform better when tractable. Score-based methods quickly become intractable because the number of DAGs to be scored grows super-exponentially with the number of nodes. The number $N_{\text{DAG}}(n)$ of labelled[3] DAGs with n nodes is listed in the OEIS [245] as entry A003024 for $n < 15$, at which point there are around $1.4 * 10^{36}$ possible DAGs. In general, all one can say is that it is bounded as $2^{\binom{n}{2}} \leq N_{\text{DAG}}(n) \leq 3^{\binom{n}{2}}$ (though there is a recursive formula [218]). Simply scoring each DAG separately is thus impossible, so a score-based causal discovery algorithm has to be based on an efficient scoring system.

[2] This introduces a new parameter in the cutoff, and for the purpose of this thesis, setting the cutoff at a higher fraction would be more conservative, since this leads to more colliders, which increases the size of the Markov blankets.

[3] *i.e.* no node permutation symmetry

In [85], a new way to enumerate and score DAGs was introduced, based on their topological ordering. Each graph can be assigned an ordering $\pi_{\prec}$ on its nodes by demanding that each node comes before its parents in the ordering. For example: The collider structure $A \rightarrow B \leftarrow C$ is compatible with the orderings $\pi_{\prec} = (B, A, C)$ and $\pi_{\prec'} = (B, C, A)$. The relationship between graphs and orderings is not a one-to-one map; a graph can be compatible with multiple orderings, and an ordering can be compatible with multiple graphs. This enumeration led to an algorithm called `order-MCMC` that constructs a Markov chain over the space of orderings, rather than the space of DAGs. It optimises the DAG structure with respect to an empirical distribution by scoring each ordering, and the effect of changing an edge (details in [85]). Scoring an order takes 2^n evaluations of the score function, which is still too slow for large n, but setting a maximum K on the number of parents a node can have reduces this complexity to $n^{(K+1)}$. To obtain an estimate for K, the authors of [141] suggest constraint-based methods like the PC algorithm. Note that MCMC sampling based on scoring orders is biased towards DAGs that appear in many orders. The empty graph, for example, is compatible with every order. A different MCMC scheme (also described in detail in [141]) called `partition-MCMC` scores DAGs differently and no longer suffers from this bias. In this thesis, however, I will only use `order-MCMC`, as it forms the basis of the main contribution of [141]: an algorithm called `iterative-MCMC`. Note that `order-MCMC` is efficient because the limit on the number of parents restricts the search space, but it is also limited in this sense. To preserve the efficiency but relax the search space, `iterative-MCMC` starts with a search space $\mathcal{H}_0$—*e.g.* the terminal state of the PC algorithm—but allows each node to have up to one extra parent outside this set. It then uses `order-MCMC` to find the *maximum a posteriori* (MAP) DAG $\mathcal{G}^*$ in $\mathcal{H}_0$, which is converted to a CPDAG. The next iteration then starts with the search space $\mathcal{H}_1 = \mathcal{H}_0 \cup \mathcal{G}^*$, where all edges in $\mathcal{G}^*$ are added to the search space. `order-MCMC` is then used again to find the MAP DAG, starting from $\mathcal{H}_1$. The algorithm terminates when $\mathcal{H}_{i+1} = \mathcal{H}_i$, *i.e.* when adding parents no longer results in a better DAG. If the search space grows too much, the algorithm will consume too many resources, so the authors suggest imposing a limit on the maximum number of parents any node can have. As long as the final search space does not include any nodes that have reached the maximum number of parents, this hyperparameter does not affect the results. Throughout this thesis, the search space is initialised with the PC algorithm. The significance threshold α is set to 0.05. The authors of [141] note that increasing α mostly increases false positives, and does little to the true positive rate, so α could be decreased if the PC algorithm takes up too many computational resources, or increased if the `iterative-MCMC` scheme needs to add too many extra edges.

The final search space contains the MAP DAG, but since a DAG can only be identified up to its Markov equivalence class, I convert this MAP estimate into its corresponding CPDAG. This MAP CPDAG is not necessarily a DAG—it can contain bidirectional edges—but it is at least as conservative as the most conservative DAG it is compatible with. By calculating the Markov blankets as $\text{MB} = \mathcal{A} + \mathcal{A}^T + \mathcal{A}^T\mathcal{A}$, where $\mathcal{A}$ is the adjacency matrix of the MAP CPDAG, one is left with Markov blankets that are at least as conservative as the most conservative Markov blankets compatible with the CPDAG. However, the MAP DAG is just one element from the final search space. Even more conservative, therefore, is to use the union of the complete final search space as the basis for the Markov blankets. Throughout this thesis, I will use the final search space as the basis for the Markov blankets, and thus the MFI estimation. The graph encoded by the adjacency matrix of this search space will be referred to as the MCMC graph. The MAP CPDAG will only be used to interpret the directional structure around interacting genes, as the CPDAG is easier to interpret causally, and sparser. I refer to the graph that defines the Markov blankets as the *quasi-causal* graph, as it does not necessarily reflect the true causal structure, and only aims to capture the conditional dependencies in the data.

In practice, I use the implementations of `order-MCMC` and `iterative-MCMC` from the `R`-package `BiDAG` ([259]). Since the gene expression is binarised, I scored DAGs using the `BiDAG` implementation of the Bayesian Dirichlet equivalent (BDe) score [105].

4.2.6 Data sets

I wanted to explore different gene expression data sets. Requirements were that they contained single-cell gene expression data, preferably generated with similar technology, on many cells ($>$ 100k to have a similar number of observations as in the RBM estimation from a previous chapter) of different types that could be separated. Since the data will be binarised, the count matrices should be relatively sparse, so a relatively low sequencing depth should not pose a problem. I chose two different mouse brain data sets—one constructed from embryonic mouse brains (Section 4.2.6.a), and one from adolescent mouse brains (Section 4.2.6.b). These two data sets contained sufficient cells, used the same library generation protocol and sequencing technology (10X *Chromium* chips, Illumina sequencing, and CellRanger demultiplexing/alignment), and both contained neurons as well as glial cells.

4.2.6.a Developmental mouse brain

The *developmental* data set corresponds to the *10X Genomics* MCD gene expression data set, already introduced in Section 2.2.5.a. In particular, I removed doublets as in Section 2.2.5.b, and applied the same QC filters as in Section 2.2.5.c. After all cells and genes were selected, the data was binarised as in Section 2.2.5.f. However, the cells and genes were annotated and selected using a different method. I used the Louvain-clustering provided by *10X Genomics*, and identified upregulated (with respect to all other cells) marker genes using the `R`-function `scran::findMarkers`. Cluster 7 had top 10 marker genes (all at FDR$< 10^{-10}$) {***Syt6***, *Gm27032*, *Slain1*, ***Pbx3***, *Rgs8*, *Fgf3*, *Nkx2-3*, *Otor*, ***Six3***, ***Myh7***}. All genes in boldface are listed on `mousebrain.org/adolescent/genes.html` [313] as markers for CNS-neurons while the other genes are not markers for any cell type (except for *Rgs8* which marks trilaminar cells). Furthermore, when calculating markers against specific other clusters, *Dlx2*, *Dlx5* and *Dlx6os1* appeared as top markers, which are known to control GABAergic neuron differentiation in developing mice [192]. Cluster 10 had top 10 marker genes (all at FDR$< 10^{-6}$) {***Gm11627***, *Abhd4*, *Mpv17*, ***Cldn10***, *Dhrs1*, *Thbs3*, *Aldoc*, *Prdx6*, *Gm20515*, *Chil1*}. All genes in boldface are listed on `mousebrain.org/adolescent/genes.html` as markers for astrocytes, while the other genes are not listed as markers for any cell type. However, *Aldoc* is a canonical astrocyte (and Purkinje cell) marker, and *Chil1* was found to be associated with astrocytes in [34]. I therefore concluded that cluster 7 is composed of neurons, and cluster 10 is composed of astrocytes. Note that I annotated

data set	N_c	N_e	μ_{UMI}	μ_b
Devel. neurons	60338	19768	0.11	0.053
Devel. astrocytes	23900	19272	0.17	0.072
Adol. neurons	20174	21094	0.14	0.066
Adol. astrocytes	19377	18702	0.051	0.030

Table 4.1: Summary of the two data sets used in this chapter. N_c: total number of cells. N_e: total number of genes expressed. μ_{UMI}: mean number of UMIs across all genes and cells. μ_b: mean expression across all genes and cells after binarisation.

embryonic mouse cells based on marker genes inferred in adolescent mice. Since the mice were late developmental (E18.5), and since there was a clear cell type signal present, the markers were deemed appropriate.

4.2.6.b Adolescent mouse brain

In addition to the developmental data set, I studied a data set of adolescent brain cells. The Mouse Brain Atlas [313] contains 160,796 QC'd type-annotated transcriptomes from various regions of the brain, taken from male and female mice at postnatal day 12 to 30 (referred to as P12-P30). I downloaded the expression data for the astrocytes and CNS-neurons from mousebrain.org/adolescent/loomfiles_level_L6.html [313] on April 26, 2021. Barcoded cDNA libraries were created using a *10X Genomics* single cell v1 kit. The authors demultiplexed and aligned all cells using the CellRanger software. Likely doublet transcriptomes were already removed by the authors [313]. To balance the cell counts of the adolescent astrocytes and neurons, only neurons from mice that were younger than 22 days (P22) were kept, which resulted in final data sets of 19,377 astrocytes and 20,174 neurons. After identifying the highly-variable genes, all expression levels were binarised as in Section 2.2.5.f.

4.2.6.c Summary of data sets

Table 4.1 shows several summary statistics of the four data sets. To balance cell counts between the adolescent and developmental data sets, I subsampled the developmental cells to a more homogeneous collection by only keeping cells from one of the two mice (mouse B). The fact that the lncRNA *Xist* is expressed in these cells indicates that mouse B was female. This resulted in more than 60k developmental neurons, and over 20k astrocytes. A similar number of genes were expressed in the four data sets, around 20k. Note the difference in the mean expression in terms of UMI counts. The adolescent astrocytes contained many fewer unique molecules than the other data sets, an effect that was still visible in the binarised data. This difference is further discussed and explored in Section 4.4.

4.3 Results

As this chapter contains the first MFI estimates on gene expression data, I started in Sections 4.3.1, 4.3.2 and 4.3.3 by quantifying how robust the estimates were with respect to the underlying data. In Sections 4.3.4 to 4.3.10 I then quantified the mechanistic content of the MFIs and validated the interactions against known biological functions and interactions of the gene products. On each data set, I calculated all 1- and 2-point interactions, but only 3-point interactions among genes that are connected on the MCMC graph, as calculating all 3-points would require in the order of 10^9 estimations. From each of the data sets, I only kept the top N most highly variable genes (before binarisation), where N will be specified in each case.

Throughout this chapter, the MFIs were compared to coexpression networks based on Pearson correlation. This was in part useful as a benchmark, to quantify how well different networks performed on the various metrics under consideration, but more importantly served as an indication of exactly what kind of structure is missed by correlation networks. The significance threshold for MFIs, based on the finite sample variance in bootstrap resamples, was set between $\alpha = 0.1$ and $\alpha = 10^{-4}$, depending on context and if large sample sizes were required. Note that any $F < 10^{-3}$ corresponds to the highest possible significance when the F-value is based on 1,000 bootstrap resampled data sets, because then $F = 0.001$ if exactly one bootstrap resample led to an estimate with a different sign.

4.3.1 Estimation requires thousands of cells and hundreds of genes

An n-point interaction can be estimated in n different ways, using the n different Markov blankets (see Section 3.2.2). An interaction is called estimable if at least one of these Markov blankets resulted in a well-defined and finite interaction, and for which all N_{bs} bootstrap resampled estimates were well-defined and finite. An interaction is called significant at level α if at least one of its possible estimates was significant at level α, *i.e.* had an F-value $\leq \alpha$. The fraction of estimable (significant) n-point interactions is $\rho_{e(s)} = N_{e(s)}/\binom{N_g}{n}$, where N_g is the number of genes, and $N_{e(s)}$ is the number of estimable (significant) interactions. As the total size of the data set increases, ρ_e is expected to asymptotically approach 1, whereas ρ_s should approach some value $0 \leq \rho_s \leq 1$, depending on the density of the network of interactions. Increasing the total number of genes included in the analysis should increase the mean size of the Markov blankets, but only up to a point, as genes are not expected to interact with arbitrarily many other genes. As Markov blankets increase, fewer interactions are estimable, and the ones that are estimable have increased variance so become less significant.

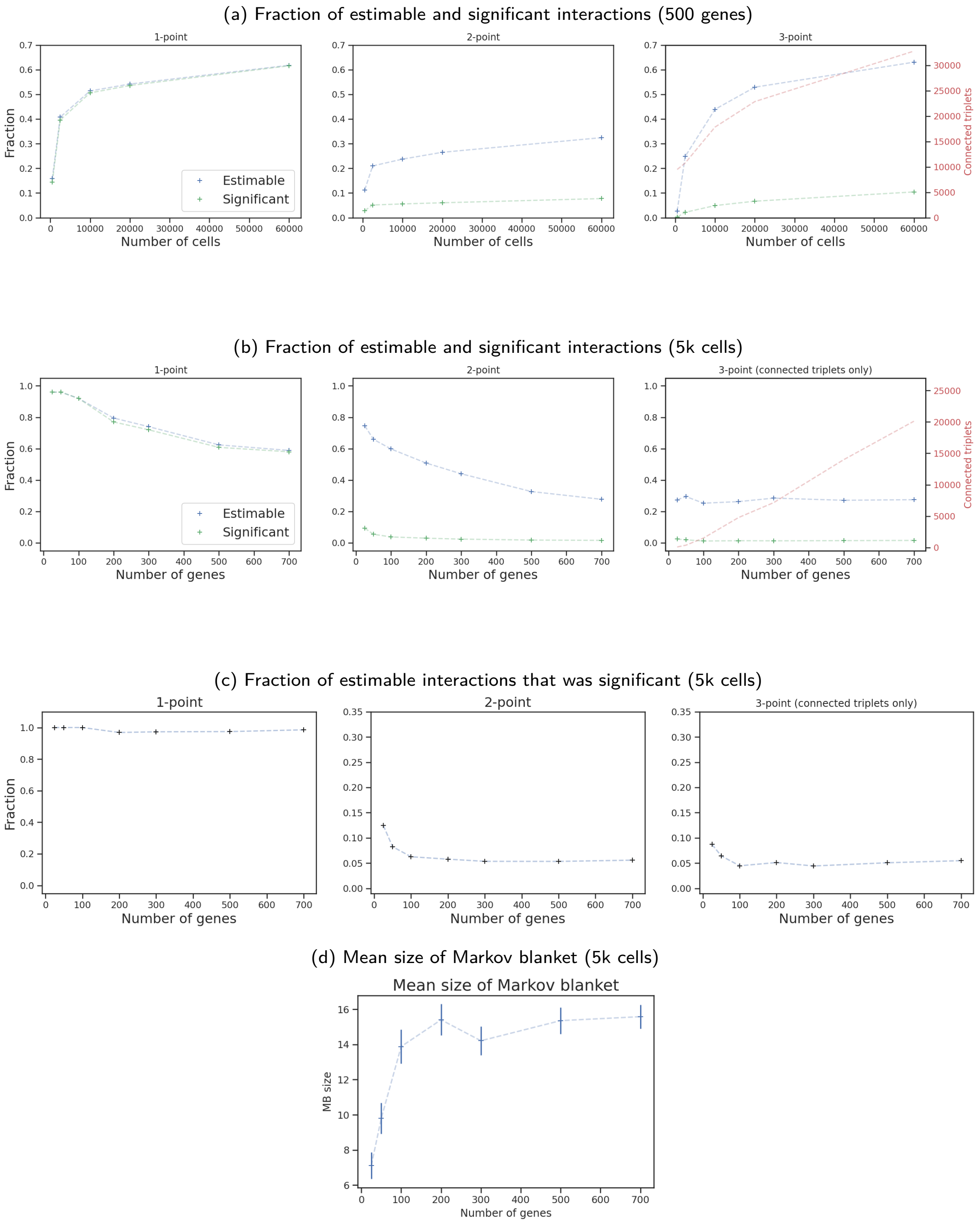
To see if these intuitions hold in practice, I calculated MFIs at up to third order in different subsets of neurons from the developmental data set, keeping only the N_g most highly variable genes. All possible 1- and 2-point interactions were calculated, but I only calculated the 3-point interactions among triplets that were connected on the MCMC graph. To get sufficiently large sample sizes across the different estimations, the significance threshold was set to $\alpha = 0.01$. Figure 4.1 shows that for a fixed number of genes, the fraction of estimable 1-, 2-, and 3-point interactions increased sharply with the number of cells at low cell number. The fractions of significant and estimable 1- and 2-point interactions both showed signs of plateauing after around 10k cells, but the fractions of 3-point interactions seem to still benefit from more data. At a fixed number of cells, increasing the number of genes N_g led to larger Markov blankets (see Figure 4.1d), which in turn decreased the fraction of estimable and significant interactions. The mean size of the Markov blankets stabilised after including around 200 genes, at which point the fraction of estimable interactions that were significant also stabilised for orders 1, 2, and 3. I therefore concluded that at least 200 highly variable genes should be included for effective estimation. Note that this only quantifies how *many* interactions are estimable/significant. Whether *individual* interactions are robust is quantified in the next sections.

4.3.2 Robustness to cell selection

If the interactions reflect biology, rather than technical noise in the data, then the interactions should be reproducible across different sets of cells from the same homogeneous population. To investigate this, I randomly selected two disjoint sets of 20,000 cells from the developmental neurons and calculated MFIs at orders 1, 2, and 3 in each of the two data sets. I considered two ways quantifying the reproducibility. Most straightforward was demanding that the 95% confidence intervals of each interaction, significant at level α, overlapped in the two data sets. This answered the question if the interaction *strength* was reproducible. However, the precise value of the estimate is less important than the sign of the significant interactions. I therefore also quantified reproducibility by the fraction of interactions that were significant at level α with the same sign in the two data sets. In Figure 4.2, the fraction of reproducible MFIs is shown, for significance thresholds of $\alpha = 1$ (all interactions), $\alpha = 0.05$, and $\alpha = 10^{-4}$ (perfectly significant). It can be seen that the reproducibility of confidence intervals was independent of the significance threshold, and worst for 1-point interactions. However, the 1-point interactions had perfectly significant sign reproducibility at all significance thresholds. The sign of the 2- and 3-point interactions became more reproducible as the estimate became more significant. This illustrates both that the significant interactions were robust to cell selection, and that the F-value is a good measure of significance and reproducibility. All these fractions were estimated on a minimum of 83 interactions (3-point interactions with $F \leq 10^{-4}$), but typically on hundreds to thousands of interactions.

4.3.3 Robustness to gene selection

Omitting dependent variables can introduce an estimation bias and lead to false positives and false negatives (as was seen previously in Sections 2.3.3.c and 3.1.1). Estimating the same interactions on a fixed set of cells, but an increasing number of genes, reveals how individual interactions depend on the selection of genes. Let $I^{(N)}$ be the value of interaction I when estimated on a total of N genes. To investigate how robust the estimate of an n-point interaction I_n is with respect to a change in N, I estimated $I_n^{(N_0)}$ in 5,000 neurons from the developmental data set. The number of genes was then increased to $N_g > N_0$, and a new estimate $I_n^{(N_g)}$ was generated. The fraction of n-point estimates $I_n^{(N_g)}$ (of which there are $\binom{N_0}{n}$) where the 95% confidence interval (CI) overlaps with the 95% confidence interval of $I^{(N_0)}$ quantifies the robustness of the estimates. The fraction of all interactions (that are estimable in both cases) for which these confidence intervals overlapped is shown in Figure 4.3a for N_0 and N_g ranging from 200 to 700. It can be seen that the interactions were robust to increasing the number of genes: 1-point interactions



(a) Fraction of estimable and significant interactions (500 genes)

(b) Fraction of estimable and significant interactions (5k cells)

(c) Fraction of estimable interactions that was significant (5k cells)

(d) Mean size of Markov blanket (5k cells)

Figure 4.1: The fraction of estimable and significant (at $\alpha = 0.01$) interactions increases with the number of cells, and decreases with the number of genes. Shown are the interactions estimated on up to 700 genes in up to 60k developmental neurons.

were the least robust, but more than 93% of the estimable 1-point interactions agreed in CI when increasing the number of genes from 200 to 700. Under the same increase in the number of genes, more than 99% of estimable 2-point interactions agreed. The 3-point interactions seemed perfectly robust, but this could be attributed to the fact that they had a larger variance and were harder to estimate, so there were fewer to compare. For downstream analysis, the main quantity of interest is the sign of the significant

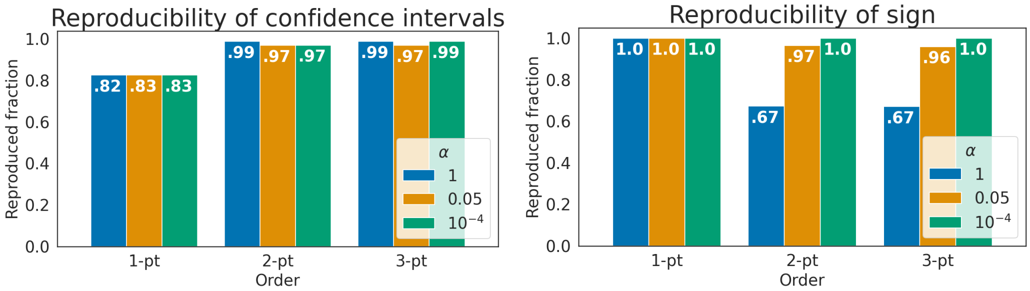
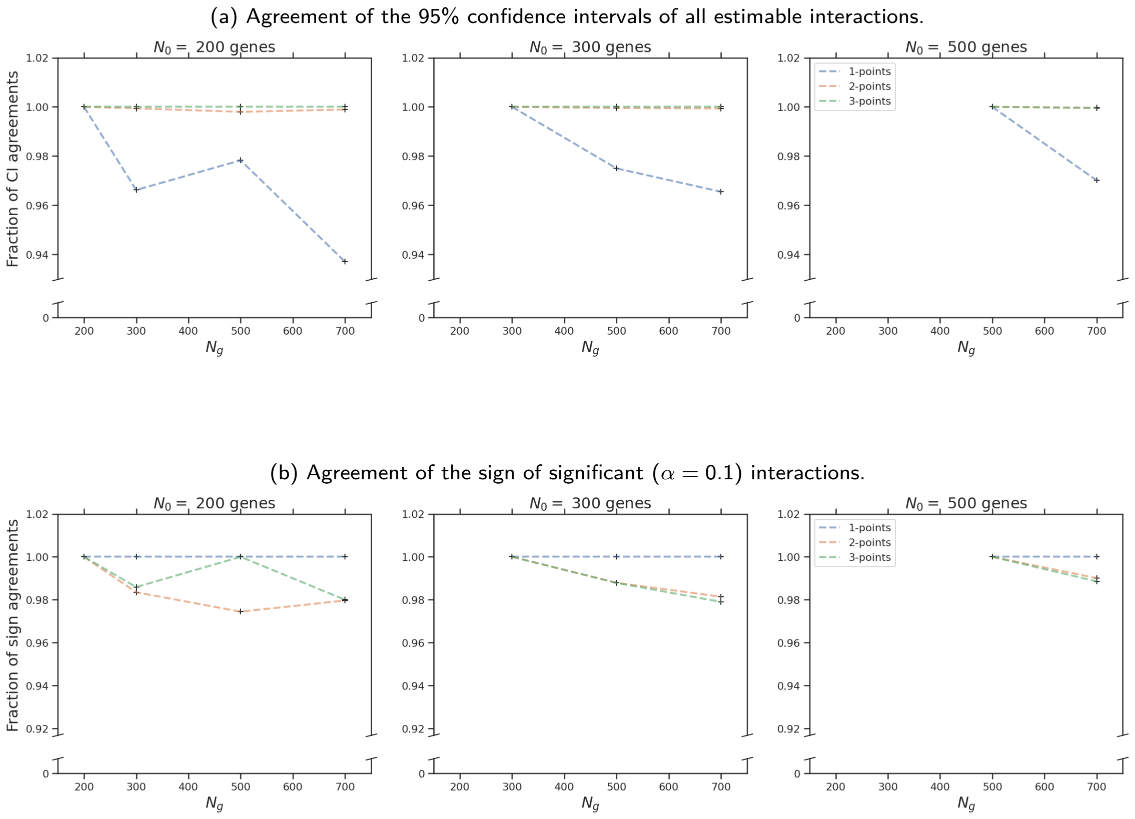


Figure 4.2: Reproducibility across two similar data sets (of 20k cells and 500 genes) of the 95% confidence intervals was high for the 2- and 3-point interactions, while even perfectly significant 1-point interactions had overlapping confidence intervals in only 83% of cases. However, all 1-point interactions agreed in sign, while only the significant 2- and 3-point interactions were reproducible.

interactions. Figure 4.3b shows the fraction of significant ($\alpha = 0.1$) interactions that differed in sign when re-estimating the interaction on more genes. Note that a larger α is more conservative in this case. As N_0 increased, a larger fraction of signs agreed upon increasing N_g. More than 95% of significant 3-point interactions kept their sign and significance upon increasing the number of genes from 200 to 700. From this, I concluded that when more than the 200 most highly variable genes were included, the interactions up to order 3 were robust with respect to gene selection.


(a) Agreement of the 95% confidence intervals of all estimable interactions.

(b) Agreement of the sign of significant ($\alpha = 0.1$) interactions.

Figure 4.3: Upon increasing the number of genes from N_0 to N_g, agreement in confidence interval and sign stabilised after including the first few hundred highly variable genes. When increasing the number of genes from 300 to 700, all significant 1-point interactions, more than 98% of 2-point interactions, and around 96% of 3-point interactions kept the same sign.

4.3.4 Functional enrichment in 1- 2- and 3-point interactors

The results of the previous sections suggest that—at least in the developmental neurons—the MFI estimates were robust as long as a few hundred highly variable genes were included, and around 20,000 cells were used. Throughout the rest of this chapter, I therefore analysed the data sets using the top 1,000 most highly variable genes in each, keeping 19,377 cells from each data set, as this was the size of the smallest data set (the adolescent astrocytes). The significance threshold for the interactions was set to $\alpha = 0.05$.

Each order of interaction reflects a different kind of dependency, so the different orders could capture different biology. To separate the order of interaction, I selected three different sets of genes in each data set. The first set was called the *1-point interactors*, and was composed of those genes that had a 1-point interaction, but did not appear in any significant 2- or 3-point interactions. Across the data sets, there were between 243 and 497 1-point interactors. The second set was called the *2-point interactors*, and was composed of the genes that appeared in at least one significant 2-point interaction, and in no 3-point interactions (between 129 and 282 genes across the four data sets). The final set was called the *3-point interactors*, and was composed of the genes that

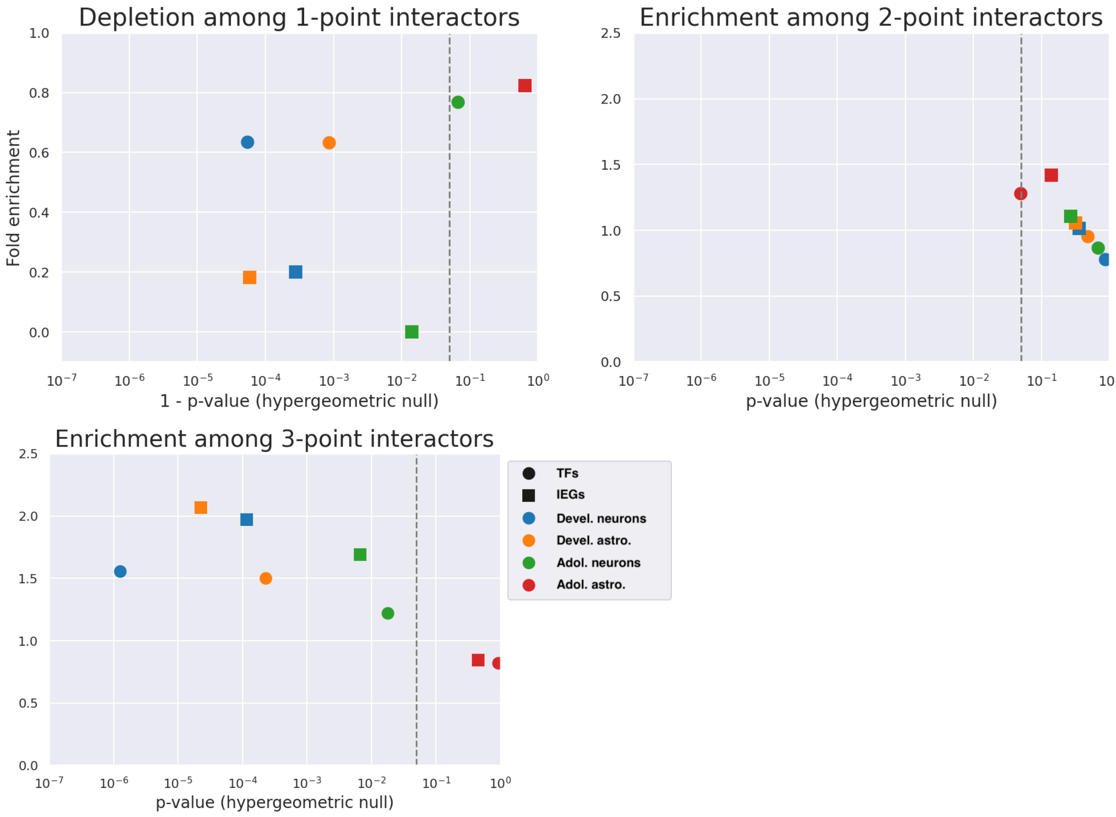


Figure 4.4: Genes with only 1-point interactions were depleted in both transcription factors (TFs, circles) and immediate-early genes (IEGs, squares), while the 3-point interactors were enriched in these genes. This held in three out of four data sets. The 2-point interactors showed no clear enrichment in either of the two classes of gene. The dashed line indicates the $p = 0.05$ threshold.

appeared in at least one significant 3-point interaction (between 369 and 475 genes across the four data sets). To make these sets of comparable size across data sets, the significance threshold was set at $\alpha = 0.05$ so that all gene sets contained between 129 and 497 genes. Given a reference list of genes, I then calculated enrichment in transcription factors (TFs) and immediate-early genes (IEGs)—two important classes of regulatory genes—with respect to a hypergeometric null hypothesis. For the reference TFs, I used the 1,623 mouse transcription factors from [116]. For the reference IEGs, I took the union across the references [275, 277, 279], leading to 108 genes. Enrichment in these two data sets is shown in Figure 4.4. It can be seen that in all data sets—except for the adolescent astrocytes—the 1-point interactors were depleted in IEGs, and to a lesser extent also in TFs. This serves as evidence supporting the hypothesis that non-interacting genes do not primarily regulate other genes. The 2-point interactors were neither enriched nor depleted in TFs or IEGs, while the 3-point interactors—except those from adolescent astrocytes—showed enrichment in both classes of genes. That 3-point interactors are enriched in these regulatory genes serves as evidence supporting the hypothesis that genes that do interact play a regulatory role in the cell.

A more agnostic approach is to query external databases using `gProfiler` for enrichment in functional gene annotations. I only report here on the 3-point interactors, as

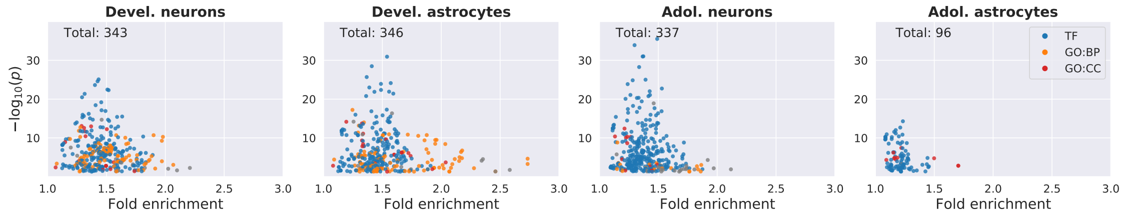


Figure 4.5: Functional enrichment among the genes that appear in a 2- or 3-point interaction. Almost all significant enrichments ($\alpha_{\text{g:SCS}} = 0.05$) were transcription factor binding motifs from the Transfac (TF), or gene ontology annotations from the biological process (GO:BP) or cellular component (GO:CC) graphs. Grey markers correspond to terms from any of the other databases.

the 2-point interactors show only weak or few enrichments. Across the data sets, the most significant enrichment results for the 3-point interactors were transcription factor binding motifs in cis-regulatory elements from the Transfac database, or gene ontology annotations from the biological process or cellular component graphs. The fold enrichment and significance are shown in Figure 4.5. It can immediately be seen that all data sets show significant but low-fold enrichment for specific transcription factor binding motifs (in blue). In all four data sets, the top 12 most significant enrichments ($10^{-37} < p_{\text{g:SCS}} < 10^{-10}$) only contain motifs for the five transcription factors *Zf5* (*Zbtb14*), *Foxn4*, *Kaiso* (*Zbtb33*), *Ben* (*Gtf2ird1*), and *E2f* (the whole family, but also *E2f-1* individually).

Also visible in Figure 4.5 is that while the most significant enrichments are TF binding motifs, the enrichment is not very strong, roughly between 1.1 and 1.6-fold. The strongest enrichments tend to be gene ontology terms (in orange and red). Sorted by fold enrichment, the 15 strongest enrichments are listed in Table 4.2. The developmental data sets show a clear development signal. The 3-point interactions in the developmental neurons correspond to the general development of the brain (the *Krox* gene family includes *Krox20*/*Egr2*, which is essential in the development of the hindbrain in mice [261]). The two strongest terms in the developmental astrocyte 3-point interactions show the negative regulation of neuron development, which reflects the shared but diverging lineages of neurons and astrocytes. The adolescent data shows a more diverse pattern. The neurons show more complex phenotypes at the level of the organism, mostly related to the functioning of the nervous system, rather than its development. Beyond the top 15 reported in Table 4.2, other Human Phenotype ontology terms significantly ($p_{\text{g:SCS}} < 0.05$, $\geq$ 1.5-fold) associated with these genes include *Abnormality of higher mental function*, *Neurodevelopmental delay*, *Hypertonia* and *Intellectual disability*. These neurological phenotypes are not significant in the astrocytes, which show much weaker enrichment in general, and mostly in GC-rich binding motifs for TFs from the *SP*/*KLF*-gene family, which is known to regulate processes like tissue differentiation, cell growth, and embryonic development [129, 199].

Comparing the enriched terms between the developmental and the adolescent data shows the nature of the data sets reflected in the 3-point interactions: the developmental data set led to interactions among genes that steer development, while the adolescent data led to interactions relevant to the functioning of the fully developed nervous system.

Developmental Neurons

Source	ID	Term	$p_{\text{g:SCS}}$-value	Fold
GO:BP	GO:0021537	telencephalon development	0.003	2.1
GO:MF	GO:0008134	transcription factor binding	0.017	2.0
GO:BP	GO:0060322	head development	3.6e-11	2.0
GO:BP	GO:0007420	brain development	3.8e-10	2.0
GO:BP	GO:0030900	forebrain development	0.00017	2.0
GO:BP	GO:0007409	axonogenesis	0.0056	2.0
GO:BP	GO:0032990	cell part morphogenesis	3.1e-05	1.9
GO:BP	GO:0032989	cellular component morphogenesis	8e-06	1.9
GO:BP	GO:0048858	cell projection morphogenesis	6.5e-05	1.9
TF	TF:M00982_1	Factor: KROX; motif: CCCGCCCCCRCCCC; match cla...	2e-06	1.9
GO:BP	GO:0007417	central nervous system development	7.1e-11	1.9
GO:BP	GO:0120039	plasma membrane bounded cell projection morpho...	0.00028	1.9
GO:BP	GO:0048812	neuron projection morphogenesis	0.00056	1.9
GO:BP	GO:0061564	axon development	0.016	1.9
TF	TF:M05599_1	Factor: WT1; motif: NGCGGGGGGGTSMMCYN; match c...	7.7e-05	1.9

Developmental Astrocytes

Source	ID	Term	$p_{\text{g:SCS}}$-value	Fold
GO:BP	GO:0031345	negative regulation of cell projection organiz...	2.3e-05	2.7
GO:BP	GO:0010977	negative regulation of neuron projection devel...	0.00053	2.7
GO:BP	GO:0006260	DNA replication	0.044	2.5
REAC	REAC:R-MMU-1640170	Cell Cycle	0.00012	2.3
REAC	REAC:R-MMU-69278	Cell Cycle, Mitotic	0.0003	2.3
GO:BP	GO:0051493	regulation of cytoskeleton organization	0.015	2.3
GO:BP	GO:0031344	regulation of cell projection organization	4.3e-07	2.2
GO:BP	GO:0010975	regulation of neuron projection development	4e-05	2.2
GO:BP	GO:0051960	regulation of nervous system development	4e-05	2.2
GO:BP	GO:0050767	regulation of neurogenesis	0.00037	2.2
GO:BP	GO:0120035	regulation of plasma membrane bounded cell pro...	1e-06	2.2
GO:BP	GO:0051129	negative regulation of cellular component orga...	0.00027	2.1
GO:BP	GO:0061564	axon development	0.0013	2.1
GO:BP	GO:0006974	cellular response to DNA damage stimulus	0.047	2.1
GO:BP	GO:0007409	axonogenesis	0.012	2.1

Adolescent Neurons

Source	ID	Term	$p_{\text{g:SCS}}$-value	Fold
MIRNA	MIRNA:mmu-let-7b-5p	mmu-let-7b-5p	0.011	2.1
HP	HP:0000729	Autistic behavior	0.00066	2.0
HP	HP:0002079	Hypoplasia of the corpus callosum	0.024	2.0
MIRNA	MIRNA:mmu-miR-340-5p	mmu-miR-340-5p	4e-05	1.9
GO:BP	GO:0048667	cell morphogenesis involved in neuron differen...	0.0013	1.9
TF	TF:M04506_1	Factor: Egr1; motif: NNMCGCCCMCTCAMWN; match c...	0.01	1.9
GO:BP	GO:0061564	axon development	0.019	1.9
HP	HP:0007367	Atrophy/Degeneration affecting the central ner...	0.021	1.8
HP	HP:0002538	Abnormal cerebral cortex morphology	0.04	1.8
HP	HP:0001257	Spasticity	0.011	1.8
HP	HP:0000486	Strabismus	0.00083	1.8
HP	HP:0000549	Abnormal conjugate eye movement	0.00083	1.8
GO:BP	GO:0000904	cell morphogenesis involved in differentiation	0.038	1.8
HP	HP:0004305	Involuntary movements	0.00078	1.8
TF	TF:M03814_1	Factor: BTEB2; motif: GNAGGGGGNGGGSSNN; match ...	0.01	1.8

Adolescent Astrocytes

Source	ID	Term	$p_{\text{g:SCS}}$-value	Fold
GO:CC	GO:0031975	envelope	0.0029	1.7
GO:CC	GO:0031967	organelle envelope	0.0029	1.7
GO:CC	GO:0031090	organelle membrane	4.7e-05	1.5
TF	TF:M01273	Factor: SP4; motif: SCCCCGCCCCS	0.0019	1.5
TF	TF:M05547_1	Factor: ZAC; motif: KGGGCCR; match class: 1	0.0022	1.5
TF	TF:M09692_1	Factor: GKLF; motif: WGGGYGKGGCCN; match class: 1	2e-05	1.4
TF	TF:M01588_1	Factor: GKLF; motif: GCCMCRCCCNNN; match class: 1	0.00053	1.4
TF	TF:M10278_1	Factor: KLF3; motif: NNNNNNGGGCGGGGCNNGN; matc...	0.0011	1.4
TF	TF:M01783_1	Factor: SP2; motif: GGGCGGGAC; match class: 1	0.031	1.4
TF	TF:M00243	Factor: Egr-1; motif: WTGCGTGGGCGK	0.045	1.3
TF	TF:M07395_1	Factor: Sp1; motif: NGGGGCGGGGN; match class: 1	0.0034	1.3
TF	TF:M01858	Factor: AP-2beta; motif: GCNNNGGSCNGVGGGN	0.025	1.3
TF	TF:M03567	Factor: Sp2; motif: NYSGCCCCGCCCCCY	5.4e-05	1.3
TF	TF:M05547	Factor: ZAC; motif: KGGGCCR	2.6e-05	1.3
TF	TF:M08867	Factor: AP2; motif: GCCYGSGGSN	1.1e-05	1.3

Table 4.2: The top 15 most strongly enriched terms among the 3-point interactors in the four data sets. The developmental data sets showed a clear developmental pattern, while the adolescent astrocytes sets showed a more mature set of terms. The *p*-value is g:SCS corrected. Note that the adolescent astrocytes were enriched in a term from the MIRNA source, which corresponds to the miRTarBase [117] which annotates targets of micro RNAs: small (∼ 22 nucleotides) single-stranded, noncoding RNA molecules that negatively regulate target mRNAs.

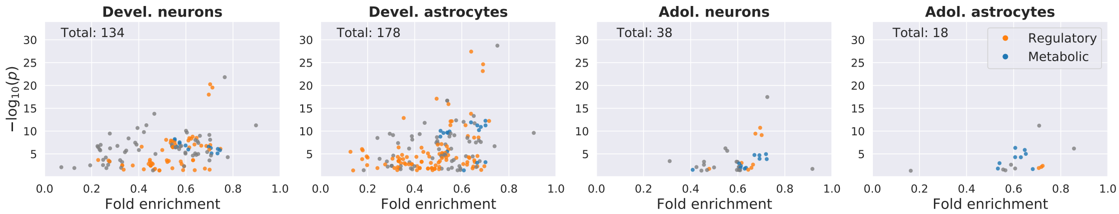


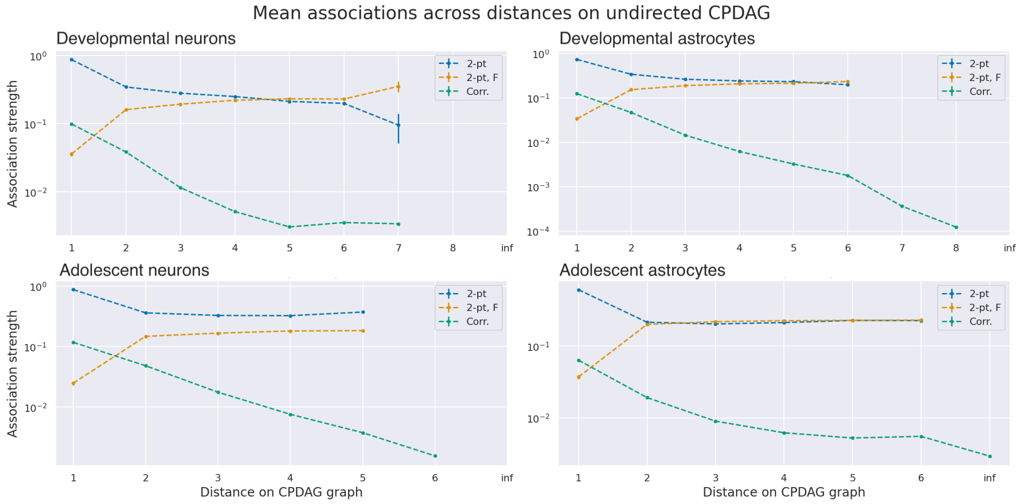
Figure 4.6: Significance and fold enrichment of GO:BP terms in a depletion query for the 1-point interactors. In all four data sets, the pure 1-point interactors were depleted in regulatory and metabolic processes. Regulatory terms were defined as containing the strings ‘regul’ or ‘respon’, and metabolic terms as containing ‘metab’ or ‘olic’. It was then manually verified that this led to correct annotations. Grey markers correspond to terms not containing any of these strings.

Testing for depletion rather than enrichment gives the inverse view on the interactions. Querying the same databases resulted in no significant depletions for the 2- or 3-point interactors, but the 1-point interactors were depleted in many transcription factor binding motifs, and regulatory and metabolic processes (see Figure 4.6 for an overview of the biological processes).

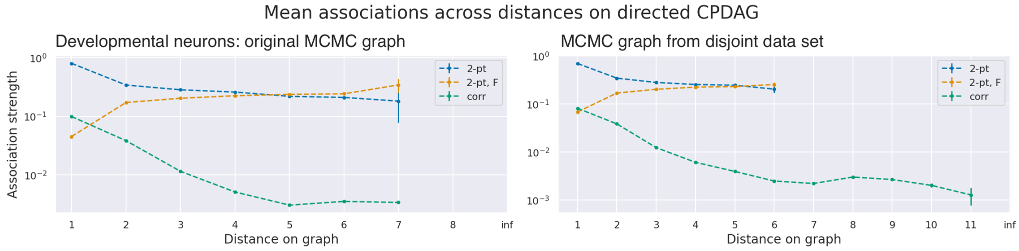
4.3.5 MFIs separate direct from indirect causal effects

Unconditioned correlations can be transitive: if A is correlated with B, and B is correlated with C, then this can induce a correlation between A and C that does not reflect a direct causal connection. Such transitive correlations thus link causally distal genes. A measure of causal proximity between two genes is the distance between them on a causal graph. Conditioning on the Markov blanket should make the MFIs link causally proximal genes more specifically than correlation networks do. To see if this hypothesis is supported by the data, I calculated the mean value and significance of pairwise interactions between genes that are a distance d removed on the quasi-causal CPDAG. Results are shown in Figure 4.7. It can be seen that across the four data sets, the interactions and their significances change most between a distance of 1 and 2, while the Pearson correlation decays at a constant rate over a much larger distance. This shows that the pairwise interactions can indeed separate direct from indirect effects. To calculate the distances, the directionality in the graph has been ignored, since parents in a pure collider should be able to interact, but could be unreachable on the directed graph, while being only a distance of 2 apart on the undirected graph.

One might worry that this is a reflection of *double-dipping*, *i.e.* using the same data twice. The CPDAG is indeed estimated using the same data as was used for the estimation of the interactions. To rule out the possibility of double-dipping distorting the conclusion, a similar CPDAG was estimated on a completely unseen set of developmental neurons. Figure 4.7b shows that the effect became slightly weaker, but the difference between the Pearson correlation and the interaction nevertheless remains clearly visible.



(a) Mean association strength (2-point interaction, its significance, or Pearson correlation) of pairs that are a distance d apart, as a function of d. Correlations (in green) decayed at a constant rate across a range of d (roughly up to $d = 5$ or $d = 6$), while the coupling strengths (in blue) of the MFIs and their significances (in orange) mostly decayed between $d = 1$ and $d = 2$, and stayed roughly constant beyond that. Error bars are the standard error on the mean.



(b) Even when estimating the quasi-causal graph and the interactions on disjoint data sets from the same homogeneous population, the difference in causal proximity between correlating and interacting genes remained clearly visible.

Figure 4.7: The MFIs disentangle direct and indirect effects, as measure by distance on quasi-causal graphs.

4.3.6 Significant interactions increase semantic similarity

Genes whose products interact might be implicated in some of the same biological processes, or colocalise inside the cell. To investigate this, I calculated the semantic similarity between interacting pairs of genes. In Figure 4.8, the semantic similarity is shown for pairs that share a significant 2-, or 3-point interaction (at level $\alpha = 0.05$), or no interaction beyond first order at all. For comparison, the mean semantic similarity across all pairs (regardless of interaction) is also shown, as well as the top N correlating pairs, where N is the total number of pairs with a significant 2-point interaction. Across the three different ontologies, a significant interaction increased semantic similarity. The 3-point interactors consistently outperformed the non-interactors in all ontologies, but most strongly so in the developmental data sets. However, correlating genes were semantically as similar as genes that shared a 3-point interaction. The 2-point interactors were indistinguishable from non-interactors in terms of GO:BP similarity but performed well in the other two ontologies. These 2-point interactors thus seemed to be local, involving similar molecules, but did not correspond to any particular biological process or pathway. This pattern was particularly strong in the developmental data sets, less so for the adolescent neurons, and absent in the adolescent astrocytes, which showed no clear pattern.

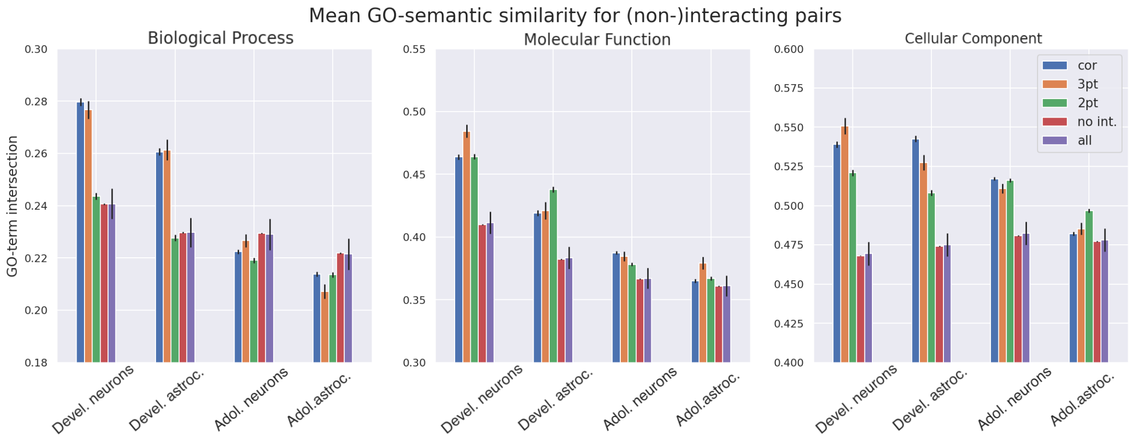


Figure 4.8: Shown are the mean and standard error on the mean of the semantic similarity of gene pairs that were linked together by either significant 2- or 3-point interactions, no interaction, or a Pearson correlation. In almost all cases, significant 3-point interactions resulted in the most semantic similarity, while pairs that did not interact were as semantically dissimilar as any two random genes.

4.3.7 3-point interacting colliders are enriched in transcription factors

After estimating the graph of conditional dependencies, triplets of genes that show a collider structure $A \to B \leftarrow C$ can be identified, indicating that the data is best explained by assuming that the causal effects flow from genes A and C to B. To see if this causal structure agrees with biology, I calculated enrichment in transcription factors of parents in such collider triplets (genes A and C), relative to the children (gene B), since the expression of TFs should be causally upstream of the expression of their targets. Note that bidirectional edges can lead to a triplet containing multiple colliders. TFs are not always only upstream—they can be downstream of other TFs—but should be so more often than randomly chosen genes. I calculated this relative enrichment in all collider triplets in the MCMC graph, which is denser than the MAP CPDAG so contains more colliders. In Figure 4.9, it can be seen that in all data sets, the upstream parent genes were enriched in transcription factors relative to their downstream children genes. Moreover, the enrichment increased when considering only colliders with a significant 3-point interaction. This simultaneously shows three things. First, it shows that the orientation induced by the quasi-causal graph was consistent with the biology: transcription factors tend to be upstream—their targets downstream. It further confirms that MFIs captured regulation by transcription factors, and indicates that the regulatory relationships among transcription factors and their targets are—at least in part—best described as a higher-order dependency. Note that this does not imply that the 3-point MFIs capture causal TF-target binding interactions, since a 3-point interaction could just as well be a reflection of a more causally distal mechanism. Instead, it supports a model in which the expression of transcription factors does not independently or additively affect the concentrations of other RNA molecules, but instead shows combinatorial effects.

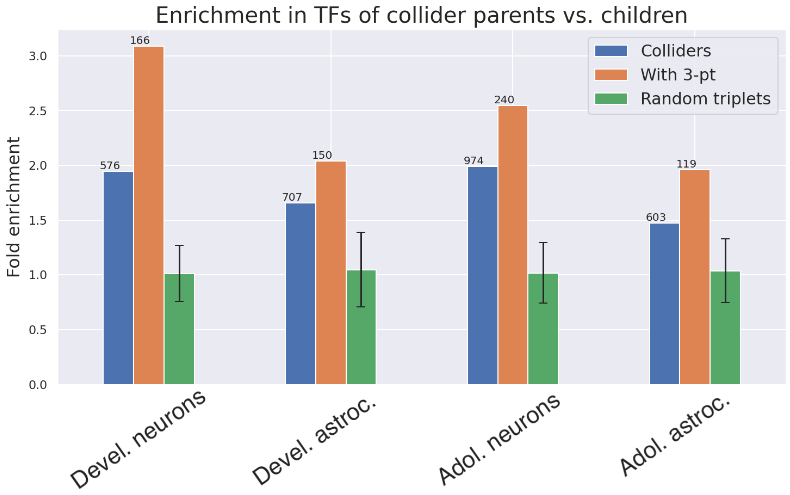


Figure 4.9: Shown is the fold enrichment in transcription factors of the parent genes in a collider triplet, relative to the children genes. A value >1 indicates that the upstream parents were more likely to be transcription factors than their downstream children genes. This enrichment is shown for all 4 data sets, both for all collider triplets and for triplets that had a significant 3-point interaction ($\alpha = 0.05$). As a control, I also selected sets of equal size but of randomly selected genes, and show the mean enrichment and the standard deviation (not the standard error on the mean) over 1,000 of these random selections. It can be seen that the parents of collider triplets were already slightly ($1.5-2$-fold) enriched in transcription factors, but this effect was stronger ($2-3$-fold) in the triplets with a significant 3-point interaction. Each bar is annotated with the number of triplets used in the calculation.

4.3.8 MFIs replicate different pathways from correlation networks.

Using the Pathway Commons database as the ground truth network, Figure 4.10 compares the ability to discover true positives of MFIs, odds ratios (ORs) and correlations. I selected all *N* significant ($\alpha = 0.05$) MFIs and odds ratios and compared these with the *N* gene pairs with the strongest Pearson correlation. At a given number of positives, the correlation network often produced more true positives than the MFIs, but this differed across the various types of interaction in Pathway Commons. Figure 4.11 shows the ratios of the number of true positives found by the MFIs and those found by Pearson correlations, separately for each category of association. MFIs performed best on interactions that regulate expression. MFIs performed worst on complex-forming associations, but well on the *Miscellaneous* category that included annotations such as *catalysis-precedes*, *neighbour-of*, *consumption-controlled-by*, *chemical-affects*, *controls-transport-of-chemical*, *controls-production-of*, *reacts-with*, and *used-to-produce*. Note that while the MFIs often produced fewer true positives, they always produced different true positives, indicating that MFIs and correlations reproduce different parts of the Pathway Commons database. The MFI's ability to recognise novel dependencies all but disappeared when not conditioning on the rest of the genes being absent, reducing the interactions to odds ratios (ORs). While the ORs performed better than correlations in most cases, a much larger fraction of their true positives overlapped with those found by correlations, relative to those found by MFIs. This large overlap cannot be explained by ORs and correlation networks both finding almost all positives, since the total number of positives in Pathway Commons is much larger than those found by either method. For the 'combined' category, for example, the total number of ground truth positives ranges from 3,996 in the adolescent astrocytes to 5,117 in the developmental astrocytes. It was thus indeed the conditioning on the Markov blanket that differentiated the MFIs from odds ratios and correlations.

A natural next question is how the true positives that MFIs uniquely identified differ from the true positives found only by correlation networks. To investigate this, I looked for GO:BP term enrichment in the genes involved in the true positives found only by correlations (the green part of the Venn diagrams in Figure 4.10) or only by 2-point MFIs (the red part of the Venn diagrams in Figure 4.10). Results are shown in Tables 4.3 and 4.4, respectively. In all data sets, except the adolescent astrocytes (that mainly show transport- and cell-migration-related enrichments), the true interactions that were only found by correlations were mainly mitotic cell-cycle related genes, whereas the true interactions found by 2-point MFIs were primarily related to development and regulation in the developmental data sets, and metabolic interactions in the adolescent data sets. Interestingly, only the MFIs in the developmental data reproduced cell-type specific interactions enriched in cell projection and synapse organisation (in which both neurons and astrocytes play a role), and brain and CNS development. Both these data sets also showed a clear $\sim$ 3-fold enrichment of the MFI-only interactions in 'cellular response to stress', which supports the finding from Section 4.3.4 that the MFIs reflect regulatory processes by IEGs. The adolescent data sets, in contrast, had their MFI-only interactions enriched in metabolism and biosynthesis-related terms.

Developmental Neurons

ID	name	$p_{\text{g:SCS}}$-value	Fold	term size
GO:0060420	regulation of heart growth	0.017	6.3	8
GO:0046620	regulation of organ growth	0.019	5.3	12
GO:0007059	chromosome segregation	1.1e-05	5.0	20
GO:0098813	nuclear chromosome segregation	0.00032	4.9	18
GO:0000819	sister chromatid segregation	0.0017	4.8	17
GO:0000070	mitotic sister chromatid segregation	0.0087	4.7	16
GO:0140014	mitotic nuclear division	0.00035	4.4	23
GO:0140694	non-membrane-bounded organelle assembly	0.035	4.1	20
GO:0014706	striated muscle tissue development	4.1e-07	3.9	40
GO:0044772	mitotic cell cycle phase transition	0.0096	3.9	24
GO:1903047	mitotic cell cycle process	1e-08	3.9	47
GO:0000280	nuclear division	0.00029	3.9	31
GO:0048738	cardiac muscle tissue development	0.041	3.8	23
GO:0048285	organelle fission	0.00014	3.8	33
GO:0007517	muscle organ development	0.01	3.7	27

Developmental Astrocytes

ID	name	$p_{\text{g:SCS}}$-value	Fold	term size
GO:0006260	DNA replication	1.3e-08	4.9	20
GO:0006281	DNA repair	3.4e-06	4.4	21
GO:0000819	sister chromatid segregation	0.00034	4.3	18
GO:0000070	mitotic sister chromatid segregation	0.03	4.1	15
GO:0007059	chromosome segregation	3e-06	4.1	25
GO:0098813	nuclear chromosome segregation	0.0011	3.9	21
GO:1901990	regulation of mitotic cell cycle phase transition	0.0045	3.8	20
GO:0051301	cell division	2.5e-07	3.8	34
GO:0044772	mitotic cell cycle phase transition	0.00057	3.7	25
GO:1903047	mitotic cell cycle process	2.9e-12	3.6	54
GO:0006974	cellular response to DNA damage stimulus	4.9e-07	3.6	37
GO:0045786	negative regulation of cell cycle	0.0089	3.6	23
GO:0140014	mitotic nuclear division	0.0015	3.5	26
GO:0000278	mitotic cell cycle	5.6e-15	3.4	70
GO:0045787	positive regulation of cell cycle	0.014	3.3	26

Adolescent Neurons

ID	name	$p_{\text{g:SCS}}$-value	Fold	term size
GO:0007051	spindle organization	0.024	3.8	16
GO:0006511	ubiquitin-dependent protein catabolic process	0.0057	3.7	19
GO:0019941	modification-dependent protein catabolic process	0.0057	3.7	19
GO:0000070	mitotic sister chromatid segregation	0.0057	3.7	19
GO:0010498	proteasomal protein catabolic process	0.022	3.6	18
GO:0140014	mitotic nuclear division	0.00028	3.5	25
GO:0043632	modification-dependent macromolecule catabolic...	0.018	3.5	20
GO:0007059	chromosome segregation	0.00023	3.4	27
GO:0000819	sister chromatid segregation	0.015	3.4	22
GO:0051603	proteolysis involved in cellular protein catab...	0.011	3.3	24
GO:0098813	nuclear chromosome segregation	0.011	3.3	24
GO:0051301	cell division	3.3e-09	3.3	48
GO:1903047	mitotic cell cycle process	5.9e-11	3.2	58
GO:0000278	mitotic cell cycle	1.3e-12	3.2	66
GO:0001505	regulation of neurotransmitter levels	0.028	3.2	25

Adolescent Astrocytes

ID	name	$p_{\text{g:SCS}}$-value	Fold	term size
GO:0098662	inorganic cation transmembrane transport	0.0028	3.8	29
GO:0098655	cation transmembrane transport	0.0051	3.6	33
GO:0098660	inorganic ion transmembrane transport	0.02	3.5	32
GO:0034220	ion transmembrane transport	0.0037	3.4	39
GO:0030001	metal ion transport	0.03	3.3	36
GO:0016477	cell migration	0.038	2.4	75
GO:0040011	locomotion	0.0033	2.4	93
GO:0006928	movement of cell or subcellular component	0.00018	2.4	110
GO:0010647	positive regulation of cell communication	0.035	2.2	101
GO:0023056	positive regulation of signaling	0.035	2.2	101
GO:0009987	cellular process	0.00013	1.2	729

Table 4.3: Top 15 GO:BP enrichments of the genes in true positives only found by correlation networks. The p-values are g:SCS corrected.

Developmental Neurons

ID	name	$p_{\text{g:SCS}}$-value	Fold	term size
GO:0099173	postsynapse organization	0.0036	4.4	27
GO:0014706	striated muscle tissue development	0.011	3.5	40
GO:0060537	muscle tissue development	0.011	3.4	44
GO:0061061	muscle structure development	0.0041	3.3	50
GO:0033554	cellular response to stress	0.00066	3.1	67
GO:0120035	regulation of plasma membrane bounded cell pro...	0.029	2.8	64
GO:0050808	synapse organization	0.0055	2.8	73
GO:0048858	cell projection morphogenesis	0.0076	2.8	74
GO:0032990	cell part morphogenesis	0.01	2.7	75
GO:0060322	head development	7.5e-05	2.7	103
GO:0007420	brain development	0.00034	2.7	98
GO:0032989	cellular component morphogenesis	0.015	2.6	81
GO:0007417	central nervous system development	3.2e-06	2.6	123
GO:0034330	cell junction organization	0.015	2.6	86
GO:0000902	cell morphogenesis	0.002	2.5	99

Developmental Astrocytes

ID	name	$p_{\text{g:SCS}}$-value	Fold	term size
GO:0032970	regulation of actin filament-based process	0.014	4.2	24
GO:0080135	regulation of cellular response to stress	0.039	3.4	36
GO:0050808	synapse organization	0.046	3.1	44
GO:0120035	regulation of plasma membrane bounded cell pro...	0.0024	3.0	57
GO:0120039	plasma membrane bounded cell projection morpho...	0.0057	3.0	55
GO:0048812	neuron projection morphogenesis	0.0057	3.0	55
GO:0048858	cell projection morphogenesis	0.0057	3.0	55
GO:0051129	negative regulation of cellular component orga...	0.047	3.0	48
GO:0001655	urogenital system development	0.047	3.0	48
GO:0031344	regulation of cell projection organization	0.0037	3.0	58
GO:0033554	cellular response to stress	1e-06	2.9	92
GO:0034330	cell junction organization	0.019	2.9	58
GO:0031175	neuron projection development	1.5e-05	2.9	86
GO:0032989	cellular component morphogenesis	0.0074	2.8	64
GO:0120036	plasma membrane bounded cell projection organi...	1.8e-07	2.8	105

Adolescent Neurons

ID	name	$p_{\text{g:SCS}}$-value	Fold	term size
GO:0016071	mRNA metabolic process	0.008	4.0	24
GO:0051254	positive regulation of RNA metabolic process	0.0062	2.3	97
GO:0045935	positive regulation of nucleobase-containing c...	0.012	2.2	108
GO:0016070	RNA metabolic process	2.1e-07	2.1	204
GO:0006366	transcription by RNA polymerase II	0.04	2.0	131
GO:0090304	nucleic acid metabolic process	2.8e-06	1.9	224
GO:0097659	nucleic acid-templated transcription	0.0024	1.9	172
GO:0006351	transcription, DNA-templated	0.0024	1.9	172
GO:0032774	RNA biosynthetic process	0.003	1.9	173
GO:0051173	positive regulation of nitrogen compound metab...	0.0072	1.9	167
GO:0006139	nucleobase-containing compound metabolic process	2.5e-06	1.9	239
GO:0051252	regulation of RNA metabolic process	0.0059	1.9	176
GO:0044271	cellular nitrogen compound biosynthetic process	0.00018	1.9	216
GO:0046483	heterocycle metabolic process	3.5e-06	1.9	251
GO:0018130	heterocycle biosynthetic process	0.0047	1.8	190

Adolescent Astrocytes

ID	name	$p_{\text{g:SCS}}$-value	Fold	term size
GO:0051276	chromosome organization	0.01	2.8	50
GO:0045934	negative regulation of nucleobase-containing c...	0.01	2.4	75
GO:1901575	organic substance catabolic process	0.0047	2.2	104
GO:0009056	catabolic process	0.029	2.0	127
GO:0006139	nucleobase-containing compound metabolic process	2.7e-07	1.9	246
GO:0046483	heterocycle metabolic process	2.8e-07	1.9	251
GO:0019219	regulation of nucleobase-containing compound m...	0.0033	1.9	176
GO:0018130	heterocycle biosynthetic process	0.002	1.8	183
GO:1901362	organic cyclic compound biosynthetic process	0.0021	1.8	188
GO:1901360	organic cyclic compound metabolic process	1.4e-07	1.8	268
GO:0006725	cellular aromatic compound metabolic process	3.8e-07	1.8	262
GO:0034654	nucleobase-containing compound biosynthetic pr...	0.0066	1.8	179
GO:0010556	regulation of macromolecule biosynthetic process	0.0057	1.8	183
GO:0019438	aromatic compound biosynthetic process	0.005	1.8	187
GO:0009889	regulation of biosynthetic process	0.0017	1.8	201

Table 4.4: Top 15 GO:BP enrichments of the genes in true positives only found by MFIs. The *p*-values are g:SCS corrected.

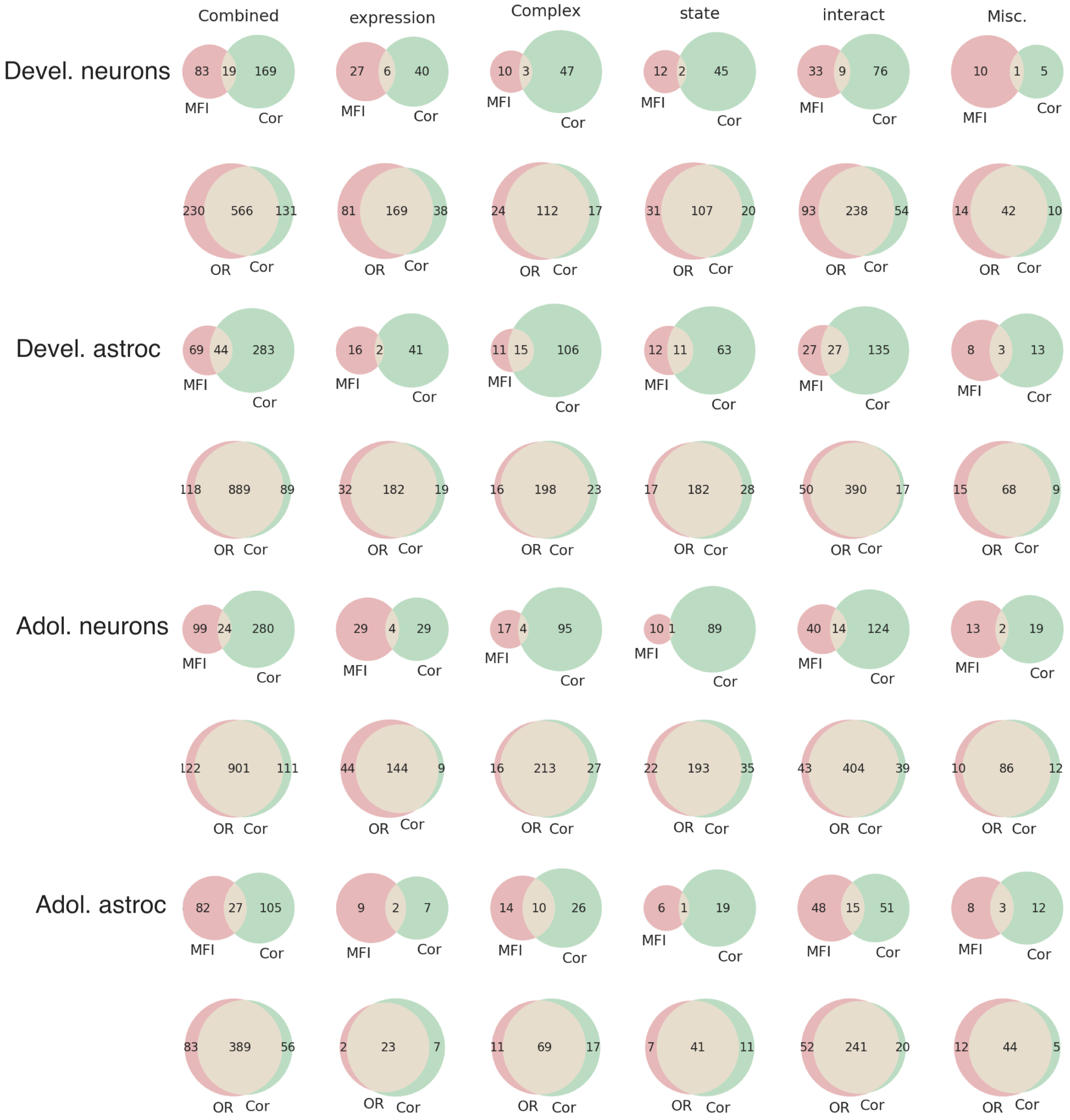


Figure 4.10: Venn diagrams of the true positives found by MFIs, log-odds ratios (OR), and Pearson correlations (Cor). Pairwise MFIs at significance level $\alpha = 0.05$ reproduced interactions from the Pathway Commons database that correlation networks did not, even when correlations generate more true positives. This effect was strongest for association that were annotated as controlling 'expression' in the Pathway Commons database. Log-odds ratios—i.e. unconditioned MFIs—reproduced mostly interactions that were already discovered by correlation networks.

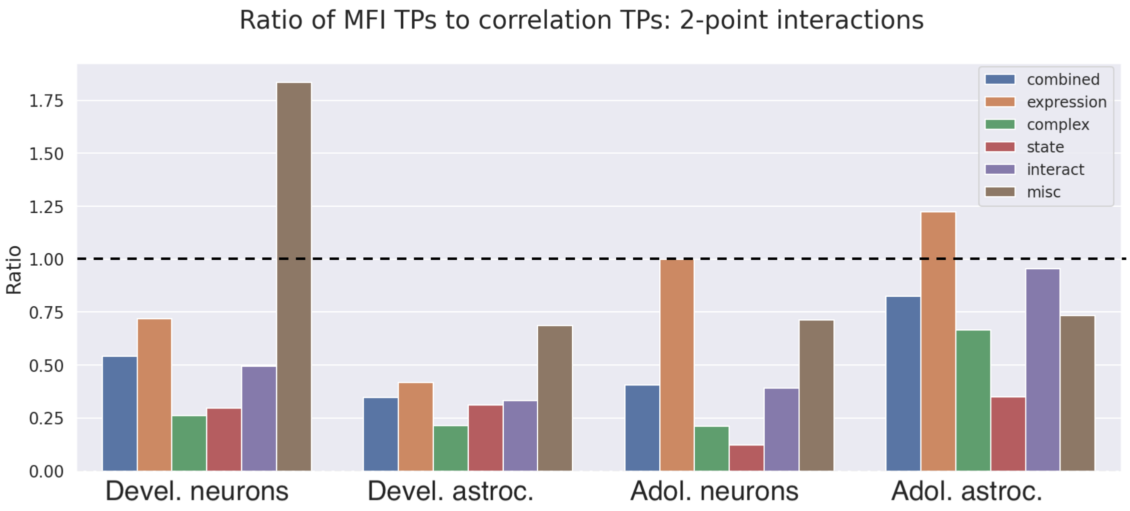


Figure 4.11: A comparison of the true positives (TPs) found by the MFIs and correlations, with a ratio of 1 dashed in black. In all four data sets, the MFIs performed best on the 'expression' category of Pathway Commons associations (ignoring the 'miscellaneous' category that combines multiple smaller categories). Among these expression regulation interactions, the MFIs from the adolescent data set performed as well as or better than correlation networks. This Figure aggregates the data from Figure 4.10.

4.3.9 3-point interactions identify genetic logic gates by sign

As shown in Section 3.3.2, all six non-trivial 2-input logic gates have a 3-point interaction. The gates OR, XOR and NAND have a negative 3-point interaction, while NOR, XNOR and AND have a positive 3-point interaction (see Table 4.5).

The converse, however, is not true: a 3-point interaction is not necessarily indicative of a logic gate. To see if there is a relationship between the 3-point interactions in the expression data and logical rules underlying the regulatory relationships in the gene expression data, I focused on the 3-point interactions of pure collider triplets (*i.e.* no bidirectional edges) since colliders are most easily interpreted as 2-input gates. I selected only those cells where the genes in the smallest of the three Markov blankets were not expressed. To identify the underlying logical relationship, I separated these cells by the 4 possible states of the parent nodes $S = \{(00), (01), (10), (11)\}$, and calculated the mean expression μ_{S_i} of the child node in each of these four partitions. Let the mean expression of the child node of the collider in these cells be μ_{out}. Each collider triplet was then represented as a binary vector of length 4, where entry i is 1 if $\mu_{S_i} \geq \mu_{\text{out}}$, and 0 otherwise. Using the output column in the truth table of a gate (using the same ordering of the four input states), each logic gate was likewise represented as a length-four vector (e.g. XOR can be represented as $(0, 1, 1, 0)$). With each triplet I associated the logic gate whose vector representation is closest to it (in terms of the L2-norm of their distance, which on binary vectors is equivalent to the Hamming distance). Table 4.6 shows that, in almost all cases, the sign of the closest gate agreed with the sign of the 3-point interaction of the triplet. The only triplets that got the 'wrong' sign—i.e. were closest to a gate with the opposite sign—corresponded to OR or AND gates, which are the more weakly coupled gates. This is further evidence that 3-point interactions can indeed be interpreted as logic-like combinatorial regulation, and that such higher-order regulation is common in developing neurons and astrocytes.

Note that I have only focused on collider triplets here. While it is reassuring that the logic gate intuition holds for this simple motif, the graph can contain more complex motifs with higher-order interactions that cannot be interpreted as logic gates but correspond to a more abstract combinatorial regulation. Finally, note that it is not surprising that the X(N)OR gates were most easily identified. Not only do these gates have the strongest 3-point coupling, but they are also independent of the underlying graph: the X(N)OR truth table is invariant under a relabelling of the in- and output nodes.

$\mathcal{G}$	$I^{\mathcal{G}}_{ABC}$
XNOR	I
XOR	$-I$
AND	$\frac{1}{2}I$
OR	$-\frac{1}{2}I$
NAND	$-\frac{1}{2}I$
NOR	$\frac{1}{2}I$

Table 4.5: The 3-point interactions for all 2-input logic gates at equal noise level are related. Repeat of Table 3.1.

Developmental		**Adolescent**	
Neurons	Astrocytes	Neurons	Astrocytes
8/8 (100%)	9/12 (75%)	42/44 (95%)	15/15 (100%)

Table 4.6: Shown is the fraction of collider triplets with a significant ($\alpha = 10^{-4}$) 3-point interactions whose sign agrees with the most similar logic gate. The sign is correctly identified in almost all cases, and for all X(N)OR gates.

4.3.10 Louvain clusters of the interaction graph reveal functional modules

Based on results from co-expression networks [202], and a general intuition of biological networks [287], it is likely that the networks of interactions are—to a certain extent—*modular*. In a modular network, the full network can be partitioned, perhaps fuzzily, into subnetworks that perform distinct biological functions, and where each partition, or *module*, interacts more strongly with nodes from the same subnetwork than with other parts of the partition. Section 4.3.4 showed that the 2-point interactions were not enriched in any gene annotations, but this could be because the network is highly modular and different modules focus on different tasks. To investigate if the network of 2-point interaction indeed shows modular functionality, I constructed the graph where each vertex is a gene, and two genes A and B are connected by an edge if and only if they shared a significant 2-point interaction at level $\alpha = 10^{-4}$. This graph was then Louvain-clustered [31] to find the clustering with optimal modularity, using the `community_multilevel` function from the `Python`-module `igraph` [61].

Figure 4.12 shows enrichment in immediate early genes (IEGs), transcription factors (TFs), and housekeeping genes (HKGs) for each of the clusters separately (showing only clusters of size larger than 1). It can be seen that there was a particularly strong enrichment in IEGs. Immediate early genes generate rapid cellular responses to a wide variety of perturbations, and it has been argued that they compute and communicate the appropriate response through combinatorial expression patterns [238, 223, 144].

Note that not only were these clusters enriched in IEGs, but they were actually predictive of IEGs. For example, cluster 36 in the developmental neurons comprised the following 32 genes: ***Atf3***, *B630019K06Rik*, *B930036N10Rik*, *Baz1a*, ***Btg2***, *Cbx3*, *Cdh13*, *Cited1*, *Cnr1*, ***Dusp1***, *Flrt2*, *Flrt3*, ***Fos***, *Hist1h1c*, *Hist1h2bc*, *Hist4h4*, *Hmgb2*, *Id1*, ***Ier2***, ***Ier3***, *Ifitm2*, ***Jun***, *Klf10*, *Neat1*, ***Nfkbia***, *Plk2*, *Prox1*, *Ptprz1*, *Sox9*, *Spry2*, *Tob1*, *Txnip*. The genes that are annotated as being IEGs in the reference list are printed in boldface, but *Flrt3* [182], *Id1* [150], *Txnip* [74], *Klf10* [255], *Plk2* (Snk) [243], *Sox9* [49], and *Spry2* [231] have all been annotated as immediate early response genes in the listed references. Including these in the reference list made developmental neuron cluster 36 more than 20-fold enriched with a p-value of 0 at machine precision. Querying these clusters for enrichment in other databases with `gProfiler`, I found concordant biology. For example, the same cluster 36 was enriched in the GO:BP terms listed in Table 4.7, which showed a clear apoptotic response signature. That neuronal apoptosis is regulated and mediated by IEGs has been suspected for decades [106]. It thus seems that Louvain cluster 36 of the developmental neurons corresponded to a network of interactions that trigger an apoptosis response.

Source	ID	Term	$p_{\text{g:SCS}}$-value	Fold
GO:BP	GO:0042981	regulation of apoptotic process	0.016	4.4
GO:BP	GO:0043067	regulation of programmed cell death	0.016	4.4
GO:BP	GO:0010941	regulation of cell death	0.014	4.1
GO:BP	GO:0031324	negative regulation of cellular metabolic process	0.001	4.0

Table 4.7: All terms more than 4-fold enriched in cluster 36 of the developmental neurons.

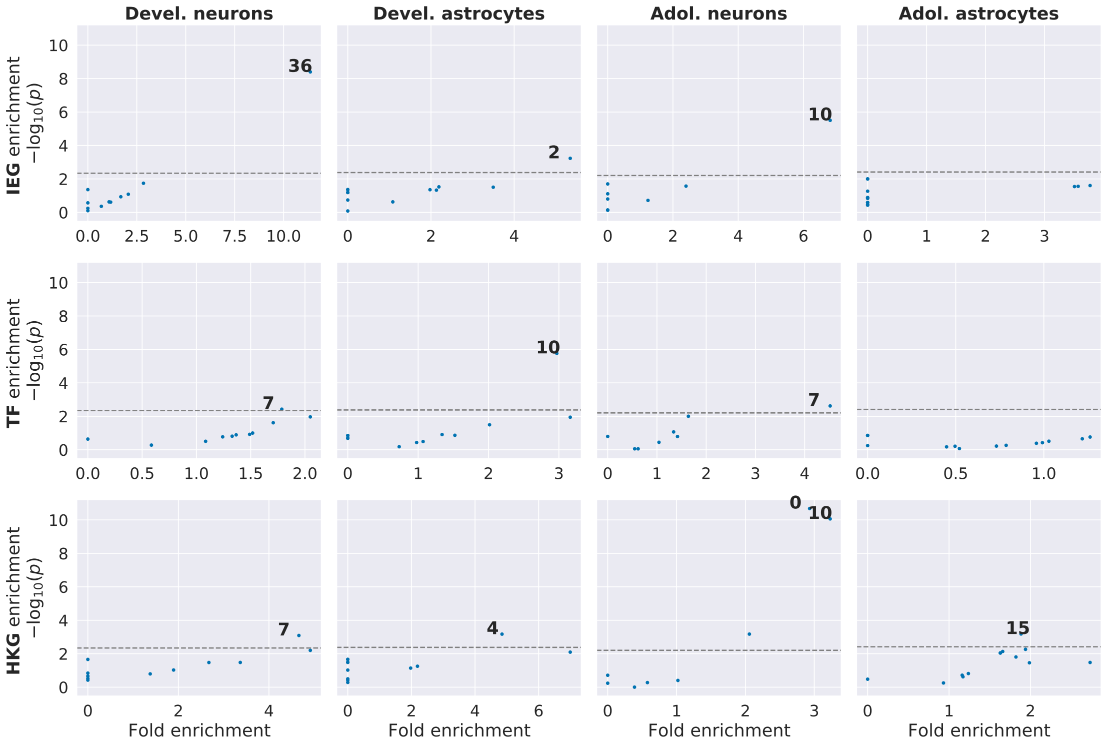


Figure 4.12: Enrichment in IEGs, TFs, and HKGs, in Louvain-clusters of 2-point interaction graphs ($\alpha = 10^{-4}$, perfect significance). Some clusters were very strongly and significantly enriched in these classes of genes, while others were not, consistent with a modular structure. Significantly and strongly enriched clusters are marked by their cluster index. The dashed line indicates the Bonferroni-corrected p-value threshold of $\frac{0.05}{n_T}$, where n_T is the number of clusters that contain more than one gene.

In the same developmental neurons, cluster 7 was enriched in both TFs and HKGs. Querying `gProfiler`, cluster 7 was indeed also only enriched in TF binding site motifs from the Transfac database, specifically for four out of the five TFs that appeared previously in Section 4.3.4, namely *Zf5, Ben, Foxn4*, and *Kaiso*, but in addition for *Ctcf, Sp1, Irf4,* and *Irf6*. Additionally, `gProfiler` indicated that cluster 30 (not highlighted in Figure 4.12) was more than 10-fold enriched (g:SCS corrected p-value of 0.018) in the Reactome pathway `REAC:R-MMU-69275`, which corresponds to the G2/M transition. In the developmental astrocytes, cluster 4 was the most significantly enriched in HKGs, and indeed in many cell-cycle related terms as shown in Table 4.8.

The adolescent neuron clusters 0, 1, 2, 5, and 10 showed enrichment in binding motifs for different transcription factors in the Transfac database, but nothing else. On the adolescent astrocytes, `gProfiler` did not report any enriched terms.

Source	ID	Term	p-value	Fold
GO:BP	GO:0008608	attachment of spindle microtubules to kinetochore	0.00039	14
GO:BP	GO:0006284	base-excision repair	0.0058	14
KEGG	KEGG:03030	DNA replication	2.5e-05	14
REAC	REAC:R-MMU-69239	Synthesis of DNA	1.5e-05	12
REAC	REAC:R-MMU-69306	DNA Replication	1.5e-05	12
REAC	REAC:R-MMU-69618	Mitotic Spindle Checkpoint	0.00021	12
GO:BP	GO:0006261	DNA-templated DNA replication	0.00021	12
GO:BP	GO:0000724	double-strand break repair via homologous reco...	0.0029	12
REAC	REAC:R-MMU-141424	Amplification of signal from the kinetochores	0.0029	12
REAC	REAC:R-MMU-141444	Amplification of signal from unattached kine...	0.0029	12

Table 4.8: All terms more than 12-fold enriched in cluster 4 of the developmental astrocytes.

4.4 Discussion

In this chapter, I explored the different ways in which higher-order dependencies among genes can reveal mechanistic relationships among gene products. I did this by calculating MFIs in gene expression data from mouse neurons and astrocytes at two different stages of brain development—embryonic and adolescent—to investigate what kind of biology they correspond to. Since I focused on RNA-level expression data, I looked in particular at two classes of genes that are known to play important roles in transcriptional regulation: transcription factors (TFs) and immediate early genes (IEGs).

Across the four data sets, I found strong evidence that the different orders of interaction reflect different kinds of biology. The 1-point interactors were depleted in TFs, IEGs, many regulatory processes, and transcription factor binding sites. The 2-point interactions showed no enrichment or depletion, but the 3-point interactions showed significant enrichment in both TFs and IEGs. Genes with 3-point interactions were also enriched in transcription factor binding sites, most strongly so for the TFs *Zf5*, *Foxn4*, *Kaiso*, *Ben*, and *E2f*. Of these, especially *E2f* stands out, as this binding motif is the main binding region for the DREAM protein complex which regulates the cell cycle, which is known to be active in the embryonic [143] and adolescent [313] mouse brain. In Section 5.3.4 of the next chapter, it will be shown that many of the 4- and 5-point interactions are composed of DREAM target genes. Furthermore, *Foxn4* is also downstream of DREAM as it is downstream from the EDM complex which comprises multiple *E2F* proteins [44] (shown in *Xenopus*). Additionally, DREAM-mediated cell cycle regulation involves the FOX gene *Foxm1*, which is known to have overlapping binding motifs with *Foxn4* [229, 173]. The other three TFs do not have a clear mechanistic interpretation, but their targets form a single module in the 2-point interaction graph, so they seem functionally related.

Even though the interactions were inferred at the level of RNA, they reflected protein function, as indicated by the fact that TF genes tended to be upstream in 3-point interacting triplets on the quasi-causal graph, in line with the role of TF proteins in transcriptional regulation. That the protein function of TFs was reflected well by the MFIs, in particular the third-order ones, is interesting in light of TF promiscuity. A limited set of TFs regulates the expression of many more target genes. This expressive power is most readily explained by some level of combinatorial regulation, *i.e.* regulation that takes into account the joint state of multiple regulatory genes, and thus naturally results in a higher-order dependency in expression data. Furthermore, it has been noted that genes and their regulatory networks can evolve independently [266], and combinatorial regulation would allow the regulatory networks to be highly modifiable and flexible. An interesting direction for further mechanistic research is thus to specifically calculate the higher-order MFIs between TFs and their targets which could give insight into the promiscuous and combinatorial nature of TF binding, both to their target sequences and to each other as homo- or heterooligomers.

The second class of genes that had a central role in the network of 3-point interactions were immediate early genes (IEGs). IEGs—being the *pathway to the genomic response*—need to respond to a vast number of inputs, model the current cell state and the incoming signal, and compute an appropriate response. To do so requires computational power or internal logic. In Chapter 3, and in Section 4.3.9, MFIs were shown to reflect logical dependencies, so the fact that IEGs played a central role in the interaction graphs, particularly of higher-order interactions, provided evidence for their computational power. A better understanding of the computational abilities of IEGs would be very valuable, not just for medicine, but also for synthetic biology, as it would open up a possibility for programmable genetic circuits (as was already anticipated and explored in [212]). Note that the Boolean logic analogy is bound to be imperfect, as gene expression is not perfectly modelled as binary, and there are cis-regulatory modules that emphatically do not allow for a description in terms of Boolean logic [268]. Nevertheless, the analogy is useful in that it provides a framework for thinking about the logic of gene expression, and has been used to guide research for decades [269, 306].

One possible explanation for the central role of IEGs in the interaction networks was offered in [280], where it was shown that the dissociation step of the scRNA-seq protocol triggers the stress response of many immediate early genes. If this is the source of the IEG MFIs in the studied data sets, then that would on the one hand indicate that the MFIs indeed represent biochemical interactions, but also hide the cell-type specific biology. Alternatively, as mentioned in Section 4.1.1, a set of IEGs could have a non-zero MFI when they are simultaneously—but independently—triggered by a stimulus. Such MFIs would reflect the stimulated state of the cells rather than direct biochemical interactions.

The 2-point interactors were not enriched in anything *as a whole*, but the different Louvain-clusters corresponded to different functional modules, with an IEG module regulating apoptosis, a housekeeping gene module regulating the metaphase-anaphase transition, and a module of specific transcription factor targets. Perhaps surprisingly, it was found that genes interacting in a 2-point interaction were less semantically similar (in terms of gene-ontology annotation) than genes that were significantly correlated in the data (see Figure 4.8). This seems to conflict with the finding that the 2-point interactions separate direct from indirect effects much better than correlations (see Figure 4.7). These two findings imply that semantic similarity is not necessarily a good measure of direct association, and that even causally distal genes can share a biological annotation that is reflected in their coexpression.

Together, the validation of 1-, 2-, and 3-point interactions supports a model in which

1. A lack of interactions corresponds to a lack of functional relationships.
2. Higher-order interactions are preferentially present in regulatory functional associations.

3. Pairwise interactions are ubiquitous and not biology-specific, but do form a functionally modular network

Throughout this chapter, the conclusions were consistent in three of the four data sets, but generally weaker or absent in the adolescent astrocytes. This can be explained by looking more closely at some summary statistics of the data sets. The adolescent astrocytes had the lowest mean expression out of the four data sets (0.057 vs. 0.059, 0.074, and 0.076), the lowest mean correlation (0.009 vs. 0.012, 0.013, and 0.018), and the fewest pairs of genes that had a Pearson correlation with an absolute value stronger than 0.1 (2750 vs. 9638, 13720, and 23671). Consequently, the adolescent astrocytes had weaker, and less significant, 2- and 3-point interactions. Furthermore, after selecting the top 1,000 most highly variable genes, the developmental data sets both contained 11 housekeeping genes, the adolescent neurons 102, and the adolescent astrocytes 182, using [114] as a reference list of mouse housekeeping genes. Together, these findings indicate that the data set of adolescent astrocytes contained less dynamic gene expression, which hindered the inference of biological mechanism.

A possible future direction for further mechanistic validation would be to extend some of the methodologies from pairwise interactions to 3-point interactions. Semantic similarity is usually only calculated for pairs, but a triplet of genes can also be assigned a last common ancestor, so this would be a natural generalisation of the method (although comparisons between orders would be difficult as the last common ancestor of a triplet is bounded by the last common ancestor of the most dissimilar of the three pairs). Similarly, I only validated pairwise interactions against Pathway Commons, but this could be extended to triplets by collapsing each 3-point interaction into three pairwise interactions and validating each separately. However, if the triplet interaction was indeed separable into three pairwise interactions, then it would have been inferred as such, and thus would not be significant triplet interaction.

Another possible improvement is to extend the enrichment analyses by considering different sets of background genes. I currently used only the 1,000 highly variable genes in each data set as the background for all enrichment results, but one might argue that each of the $\sim 20,000$ genes from the original count matrices had a chance of being highly variable and thus included in the analysis, which would justify using all mouse genes as a background. Alternatively, one might argue that only those genes that are expressed in the mouse brain had a chance of being selected, and thus only expressed genes should be used as a background. This set could be based on the data set being studied, or on orthogonal mouse brain atlases like those in [147, 143]. These background sets would be easy to implement and could lead to different conclusions. The background set of the 1,000 highly variable genes I currently used is the smallest, and should therefore be the most conservative, as including more genes in the background should lead to more statistical power.

In Section 4.3.8, it was shown that it was indeed the conditioning on the Markov blanket that distinguished the MFIs from coexpression networks in their ability to recapitulate Pathway Commons interactions. Interestingly, the MFIs were *different*, but not consistently better. This might be because the confirmed interactions from Pathway Commons are often based on *in situ* or even *in vivo* experiments, which do not reproduce the situation in which the other genes in the Markov blanket are absent. To retrieve more of the gold standard interactions, or even just biologically relevant ones, it might be necessary to condition on a different state of the Markov blanket. This is a direction for future research, and will be discussed in some more detail in Chapter 6.

Finally, the graph of conditional dependencies is crucial to the estimation. I currently estimate this graph in a purely data-driven way, but one could integrate biological knowledge into the graph construction procedure by, for example, manually adding known interactions from a 'gold standard' to the PC graph before starting the MCMC optimisation step. Alternatively, all or part of the Pathway Commons database could be added to the final graph, making the Markov blankets in part biologically interpretable. Since adding edges can only increase the variance of the estimates, not the bias, this seems worth exploring. It would be especially interesting to see if this could improve the MFIs ability to reproduce interactions from the database that was used to augment the quasi-causal graph, since that would allow the estimation to be *fine-tuned* to a particular goal. However, drawing the quasi-causal graph from multiple sources makes the mathematical interpretation of the estimated interactions more difficult, so should probably only be done after extensive validation.

In conclusion, this chapter offered insight into higher-order associations in gene regulation and provided multiple starting points for further research into the mechanisms. The results suggest that MFIs are a useful tool for studying higher-order interactions in gene regulation, and that they can be used to infer functional modules in gene regulatory networks. However, no strong evidence was found that MFIs recapitulate known pathways directly and with a precision beyond that of coexpression networks. Moreover, there was no clear way to interpret or validate the causal or mechanistic content of higher-order interactions, as they are fundamentally model-free and undirected. Therefore, I shifted the research direction towards using the MFIs for cell state inference, which will be the subject of the next chapter.

Chapter 5

Higher-order cell states and types

Everything studied in Biology represents a system of systems; itself part of a higher-order system.

François Jacob [122] (translation mine)

5.1 Introduction

5.1.1 Interactions as cell type or state

As mentioned in Section 1.3, cell states and types are usually defined by clustering cells in expression space and then assigning a biological identity to the clusters through their respective *marker-genes*: those genes that are *differentially* (usually more highly) expressed in the cluster of interest relative to other clusters. While a cluster can have multiple marker-genes, each gene marks the cluster independently, and possibly redundantly. Such clusterings have historically proven to reveal the various cell types present in a population, yet they are not compatible with all notions of cell identity. For example, cells of many cell types can be in the G2M transition state at the same time, but cluster with cells of their own type rather than cells in a similar mitotic phase, hiding the shared cell state. Furthermore, experiments tracing cells through development have shown that a cell fate decision is reached before cells separate in expression space [294]. This means that a differentiating cell's state changes before this can be revealed by clustering methods, and a more precise, 'pointwise' notion of cell state is necessary to describe this process.

In contrast to these expression clustering techniques that focus on differences in mean expression, a higher-order interaction reflects a pattern in the *joint* state of a number of genes, and thus reveals more complex patterns in the genes' interdependencies. This could potentially identify subpopulations by their combinatorial (*i.e.* both present and absent), joint gene expression patterns. Such combinatorial markers have been shown to improve classification of clusters in expression space in [67, 72]. Where traditional cell type annotation characterises clusters of cells by the genes that are *differentially* expressed, a 2-point interaction identifies genes that have a non-random *joint* state—they are *differentially differentially* expressed (*i.e.* the extent to which gene A is differentially expressed depends on the expression of gene B). In general, an n-point interaction could thus characterise cells by their *differentially*^{n} expressed genes. In this chapter, I assigned characteristic sets of cells to up to 5-point interactions and compared the structure found by the interactions with known biological annotations and clusterings. This represents a radically different approach to the semantics of cell state, already hinted at in the introduction in Section 1.3, where cells are identified and distinguished not by their mean difference in expression, but by a precise and higher-order dependency in their expression. Moreover, whereas conventionally each cell is categorised by cell type or state (usually from a collection of preconceived biological identities), here the various identities are first inferred from the cell population as a whole, and afterwards each state is assigned to individual cells, allowing each cell to be in multiple states at the same time. This is a biologically plausible concept, as cells can respond to multiple internal and external stimuli simultaneously, and coordinate the activity of multiple simultaneous pathways.

5.1.2 Aim and outline of this chapter

This chapter introduces the notion of characteristic MFI states, which is introduced in **Section 5.2.1**. The estimation of the interactions and states are combined into a `Nextflow` pipeline called `Stator`, which is introduced and outlined in **Section**

5.2.3. To compare the performance of `Stator` with clustering based cell type inference methods, I outlined a different data-driven approach in **Section 5.2.4**.

The structure of the MFI states is compared with data-driven clustering approaches in **Section 5.3.1**, and with expert annotations in **Section 5.3.2**. That the inferred states are robust with respect to cell sampling is shown in **Section 5.3.3**. States inferred on the neurons and astrocytes from the previous chapter are then discussed in **Section 5.3.4**, and that their semantics persist throughout a homogeneous population is shown in **Section 5.3.5**. In **Section 5.3.6**, states are inferred from a data set containing multiple cell types. An alternative method of inferring states from interactions is briefly introduced and explored in **Section 5.3.7**. Finally, **Section 5.3.8** explores the existence of beyond fifth-order interactions, before reflecting on the results from this chapter in **Section 5.4**.

5.2 Methods

5.2.1 Characteristic states of interactions

The value and significance of an interaction depends on the relative frequency of certain gene expression patterns present in the cell population. A significant interaction among n genes indicates that at least some of the joint states of the n genes occur more often in the population than expected by their marginal expression levels. For example, consider two genes A and B that are both expressed in 25% of all cells. If the two genes are independently expressed, then the probability that both are expressed in the same cell is $p(A=1,B=1)=\frac{1}{4}\times\frac{1}{4}=\frac{1}{16}$, while the probability that neither is expressed is $p(A=0,B=0)=\frac{3}{4}\times\frac{3}{4}=\frac{9}{16}$, and the probability that only one is expressed is $p(A=1,B=0)=p(A=0,B=1)=\frac{1}{4}\times\frac{3}{4}=\frac{3}{16}$. Since these two genes are independently expressed, their 2-point interaction is zero:

$$I_{AB} = \log\frac{p(A=1,B=1)p(A=0,B=0)}{p(A=1,B=0)p(A=0,B=1)} \tag{5.1}$$

$$= \log\frac{\frac{1}{16}\frac{9}{16}}{\frac{3}{16}\frac{3}{16}} = 0 \tag{5.2}$$

Deviations from this null hypothesis will result in a non-zero interaction, but might still be expected to occur simply by chance in a finite sample of independently expressed genes. For example, if the two genes are expressed in 25 out of 100 cells, the expected number of cells where both are expressed is 6.25. If in reality there are 8 cells with both genes expressed, this corresponds to a deviation of $\frac{8-6.25}{6.25}\times 100\% = 28\%$. One can associate a p-value to this deviation by recognising that the number of cells with both genes expressed, written as $\Phi_{\{1,1\}}$ is expressed binomially as $P(\Phi_{\{1,1\}}=k)=\binom{100}{k}\frac{1}{16}^{k}\frac{15}{16}^{100-k}$. Let the p-value associated to the observation of 8 cells with both genes expressed describe the probability of finding at least 8 cells with both genes expressed, under the assumption that the two genes are indeed independently expressed. This probability is given by

$$\begin{aligned} p &= P(\Phi_{\{1,1\}} \geq 8) && (5.3)\\ &= 1 - \sum_{k=0}^{7} P(\Phi_{\{1,1\}} = k) && (5.4)\\ &= 0.29 && (5.5) \end{aligned}$$

which reveals that the observed deviation is not beyond what would be expected from independently expressed genes, so the null is not rejected.

More generally, in a finite sample of N cells, the observed frequency Φ_s of a joint state $s = \{s_1, \dots, s_n\}$ of n independently expressed genes is binomially distributed as:

$$P(\Phi_s = k) = \binom{N}{k} \pi_s^k (1 - \pi_s)^{N-k} \tag{5.6}$$

where

$$\pi_s = \prod_{i=1}^{n} \left(s_i \mu_i + (1 - s_i)(1 - \mu_i) \right) \tag{5.7}$$

and μ_i is the mean expression of gene i across all cells under consideration (*i.e.* all cells in the population, or those where the Markov blanket is zero). Equation 5.6 describes the null hypothesis that the observed cell counts are the result of independently expressed genes, and gives the expected number of cells under this null: $\mathbb{E}[\Phi_s] = \pi_s N$. An observation $\Phi_s = \phi_s$ of one of the 2^n joint states of n genes can be assigned a p-value:

$$p = 1 - \sum_{k=0}^{\phi_s - 1} P(\Phi_s = k) \tag{5.8}$$

and a deviation factor[1]:

$$\left| \frac{\phi_s - \pi_s N}{\pi_s N} \right| \in [0, \infty) \tag{5.9}$$

A non-zero interaction can thus have one or more *characteristic states*: those states that significantly and positively deviate from the null hypothesis. Since a non-zero interaction reflects a higher-order dependency in the data, its characteristic states describe the gene expression patterns that are (at least partially) responsible for this dependency. The set of cells that have the n genes in that particular state—ignoring the state of the Markov blanket—form the associated set of cells. An example of a 4-point interaction and its characteristic states is shown in Figure 5.21. Throughout the rest of this thesis, I will use the term 'characteristic state' both to refer to the expression state of the n genes, and to the set of cells that are in this state. Note that these cells need not cluster in expression space: while these cells all share a particular gene expression pattern among the n genes, the expression of all other genes can be different. This makes it in principle possible to identify a cell state that does not localise in expression space.

[1] In the latest version of the method, the log 2-fold enrichment factor relative to the null is used instead of the deviation factor, and the associated one-sided p-value is corrected with the Benjamini-Yekutieli procedure.

5.2.2 Hierarchical clustering and bootstrapped dendrograms

Each characteristic state s can be represented as a binary vector $v \in \mathbb{B}^N$ of length N, where N is the total number of cells in the data, as follows: let $v_i = 1$ if the ith cell is in state s, and $v_i = 0$ otherwise. If there are K characteristic states, then the full data set can be represented as a $N \times K$ matrix, where each column represents a characteristic state, and each row represents a cell. Using this representation, the states can be hierarchically clustered by calculating the distances between the columns using a distance measure on $\mathbb{B}^N$. Note that this is emphatically not a clustering of cells based on distances in expression space—it is a clustering of states in 'cell space'. Clustering the *rows* of this $N \times K$ matrix would correspond to clustering cells in a K-dimensional state space, but since this is a more conventional clustering of cells rather than states it is not the focus of this thesis. If each interaction has a single characteristic state, clustering the states is the same as clustering the interactions. However, in general a single interaction can have multiple characteristic states, so I will refer mostly to the states directly.

A distance measure that assigns a distance $d(v, w)$ to any two states with binary representations v and w induces a hierarchy of separation among all states, which can be summarised in a dendrogram. The dendrogram can be visualised as a tree, where the leaves are the states, and the height of the branches is the distance between the states. Higher up the tree, branches contain multiple states, so distances have to be calculated not directly between states, but between branches. To do so, there are a variety of methods. For example, one could set the distance d between two branches X and Y equal to the distance between the two closest states from the branches: $d(X, Y) = \min_{v \in X, w \in Y} d(v, w)$, where v and w are binary representations of two different states. This is called *single-linkage clustering*. Alternatively, one could use the distance between the two furthest states from the branches: $d(X, Y) = \max_{v \in X, w \in Y} d(v, w)$. This is called *complete-linkage clustering*. Throughout this thesis, I used the average distance between all pairs of states from the branches: $d(X, Y) = \frac{1}{|X||Y|} \sum_{v \in X, w \in Y} d(v, w)$. This is called *average-linkage clustering*. This hierarchy by itself does not form a clustering, but it can be used to form a clustering by cutting the dendrogram at a certain height, and creating a cluster out of each branch that was cut. The height of the cut is a hyperparameter that can be tuned to obtain a specific number of clusters. I experimented with a variety of different distance measures, and the cut-off response curve for seven different distances is shown for the developmental neurons in Figure 5.1, and was similar in the other data sets. I found that the so-called Dice distance (also known as the Sørensen-Dice coefficient because it does not satisfy the triangle equality and is therefore only a *semimetric*) consistently formed dendrograms where the fewest number of branchings occurred at a distance indistinguishable from 0 or 1, which is a desirable property because it leads to the most structured dendrogram. The Dice distance between two binary vectors v and w is defined as

$$d(v, w) = \frac{2|v \wedge w|}{|v| + |w|} \tag{5.10}$$

where $|v| = \sum_i^N v_i$, so is a generalisation of the F_1-score. In summary, to go from the expression data to the states, the following steps are taken:

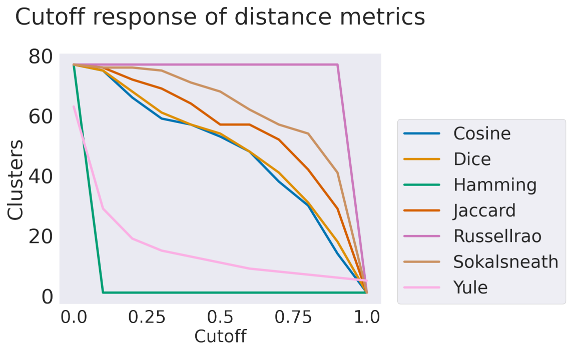


Figure 5.1: Increasing the cut-off from 0 to 1 decreased the total number of clusters at different rates for the different distance measures. The Dice distance and the cosine similarity behaved similarly, and had the desirable property that the fewest number of branchings occur at a distance close to 0 or 1. The Yule distance metric had a range beyond 1, but this is not included as it already distinguishes very few clusters at a distance of 1.

1. Calculate the interactions.
2. For each significant interaction, calculate the significantly deviating/characteristic states.
3. Represent each of these characteristic states as a Boolean vector that reflects which cell is in that state.
4. Construct the hierarchical clustering of these states using the Dice distance.
5. Cut the dendrogram at a certain height.
6. Each resulting cluster represents a final state.

To quantify how robust this hierarchical clustering is, I used two distinct bootstrapping procedures. Both approaches rely on estimating the finite sample variability by bootstrap resampling the cells to obtain N_{bs} different representations of the set of characteristic states. The first approach is based on cutting the dendrogram to obtain a particular clustering. The variability of this clustering across bootstrap resamples reflects the uncertainty associated to the finite sample, and can be quantified with the adjusted Rand index (ARI), as well as the adjusted mutual information (AMI). While the ARI is more widely used, one should in general expect unbalanced cluster sizes (since cutting a dendrogram does not necessarily result in equal cluster sizes), which is better addressed by the AMI [220]. The ARI takes values in the interval $[-1, 1]$, while the AMI falls in $[0, 1]$, both zero for random clusterings, and 1 when comparing two equivalent clusterings.

The other approach quantifies the robustness not only of the clustering at a certain threshold, but of the whole hierarchical structure. This is also done by bootstrap resampling the cells, and leads to N_{bs} different dendrograms. Each dendrogram of n leaves contains $n-1$ branchings, and thus defines $n-1$ clusters. Each of these branchings, or clusters, can be assigned a significance score based on the fraction of resampled dendrograms containing exactly that branch/cluster, called the bootstrap probability BP. However, BP is known to lead to incorrect confidence intervals (some authors call this a bias, though this terminology is disputed by [73]), so a slightly different test, called the *approximately unbiased* (AU) test, has been suggested in [241]. This test uses bootstrap resamples of different sizes in a process known as multiscale bootstrap resampling. This method has been implemented in the `R` package `pvclust` by the authors of [260], but I used a `Python` implementation from [276] to calculate both BP and AU for the dendrograms of hierarchically clustered cell states. Using this method, each individual branch can be assigned a significance, and sufficiently significant branches can be considered composite states.

5.2.3 Stator: A Nextflow pipeline to infer MFIs and states from binary data

> Computers are useless, they can only give you answers.
>
> Pablo Picasso [193]

Going from expression data to the characteristic states in section 5.2.1 involves multiple steps. The pipeline I developed to achieve this is called `Stator`, and is written in `Nextflow` [70]. `Nextflow` is a workflow management system that allows for the development of complex pipelines in a modular fashion. It is written in `Groovy`, a `Java`-based language. `Stator` is available at `github.com/AJnsm/NF_TL_pipeline`, and can be directly pulled and run from the command line, locally or on a HPC cluster using the SGE scheduler. All dependencies are packaged in various Docker containers, hosted on `hub.docker.com/u/ajnsm`, accessible to both Docker and Singularity. All python scripts run in `python 3.6`, and the R scripts in `R 4.0`. `Stator` takes in a list of parameters and data sets, calculates MFIs up to seventh order using the final search space of the MCMC graph as well as the MAP CPDAG, reconstructs the characteristic states of the significant higher-order interactions, and creates hierarchical clusterings of robust states. This is summarised in the diagram of Figure 5.2. It should be noted that the pipeline is currently still under active development, and that the documentation here provided might become outdated. The most up-to-date documentation can be found on the author's Github.

Stator starts by reading in an $N \times M$ matrix of single-cell gene expression data from N cells and M genes (or any other kind of binary observation matrix), and processing it with the `makeTrainingData.py` script using the `python`-package `scanpy`. This script can run in two different modes: `agnostic`, in which case the only preprocessing is binarising the data, and `expression`, in which case the data is normalised, log-transformed, and the top N highly variable genes are used in the analysis, where N can be set by the user. More information on the parameters and choices to be made can be found in the `README` at `github.com/AJnsm/NF_TL_pipeline/blob/develop/README.md`. After the data is prepared, the `parallelPCscript.R` script estimates the graph of conditional dependencies using the parallel, stable PC-algorithm, orienting edges using the majority rule as outlined in section 4.2.4. This PC-graph is then used as an input for the MCMC optimisation scheme in `iterMCMCscript.R`. This resulted in two quasi-causal graphs: the MAP CPDAG and the final MCMC search space. Using either of these graphs as the basis for the Markov blankets, all M 1-point interactions, and all

M^2 2-point interactions are calculated using the `estimateTLcoups.py` script. In addition, all 3-, 4-, and 5-point interactions among mutually Markov-connected genes (see Lemma 2 and Theorem 1) are calculated by the `calcHOIwithinMB.py` script. These within-Markov blanket interactions are used to guide the heuristic search for 6- and 7-point interactions by `calcHOIs_6n7pts.py`. Summary figures of all interacting tuples are then constructed (the summary of the triplet *(Id2, Id3, Slc1a3)* is shown in Figure 5.3 as an example). Among these higher-order interactors, those that have a characteristic state that deviates more than a factor of n from the Bernoulli null, at a significance level beyond θ (where n and θ are set by the user, and to 5 and 0.05 by default, respectively), are called the top deviating states and written to a file `topDeviatingHOIstates.csv`. These top deviating states are clustered in their binary representation (see Section 5.2.2), and used to draw the dendrogram of characteristic states. Finally `Stator` outputs three dendrograms of states:

1. The original dendrogram of deviating states, cut at a distance threshold set by the user.
2. The full, uncut dendrogram annotated with bootstrapped BP and AU values for each branch.
3. A dendrogram of a separate hierarchical clustering of all significant branches present in dendrogram 2. This dendrogram is not used for any of the results in this thesis, but explored in more detail in Section 6.2.2.

The pipeline offers different estimation methods in the `utilities.py` module: some use expectation values, and some use the raw probabilities. The advantage of using expectation values to estimate the MFIs is that you only need to condition on the Markov blanket of one of the interacting genes, which makes more interactions estimable. However, in practice this makes the precise value of the interaction no longer symmetric under permutation of the interacting genes. Estimating interactions using probabilities explicitly keeps the estimates symmetric.

However, perhaps a more important consideration is the difference in performance. Figure 5.4 shows the time taken by different estimation methods. It can be seen that the `numba` implementation, which uses expectation values, is by far the fastest estimation method. The speed-up is achieved by using `numba`'s just-in-time compilation, and explicitly writing out the expressions for the interactions up to 7th order. There is also a general function that can calculate the interaction at arbitrary order, but that is more than 100× slower.

An outline of the `Stator` workflow in a bash terminal on an SGE scheduler node is shown below.

Code 5.1: Stator workflow

```
$ module load singularity
$ nextflow pull AJnsm/NF_TL_pipeline
$ nextflow run AJnsm/NF_TL_pipeline -profile eddie_singularity -params-file params.json

N E X T F L O W  ~  version 22.04.3
Launching `https://github.com/AJnsm/NF_TL_pipeline` [fabulous_mahavira] DSL1 - revision: 6c5acdc435 [develop]
[f1/31129d] process > makeData (1)                      [100%] 1 of 1
```

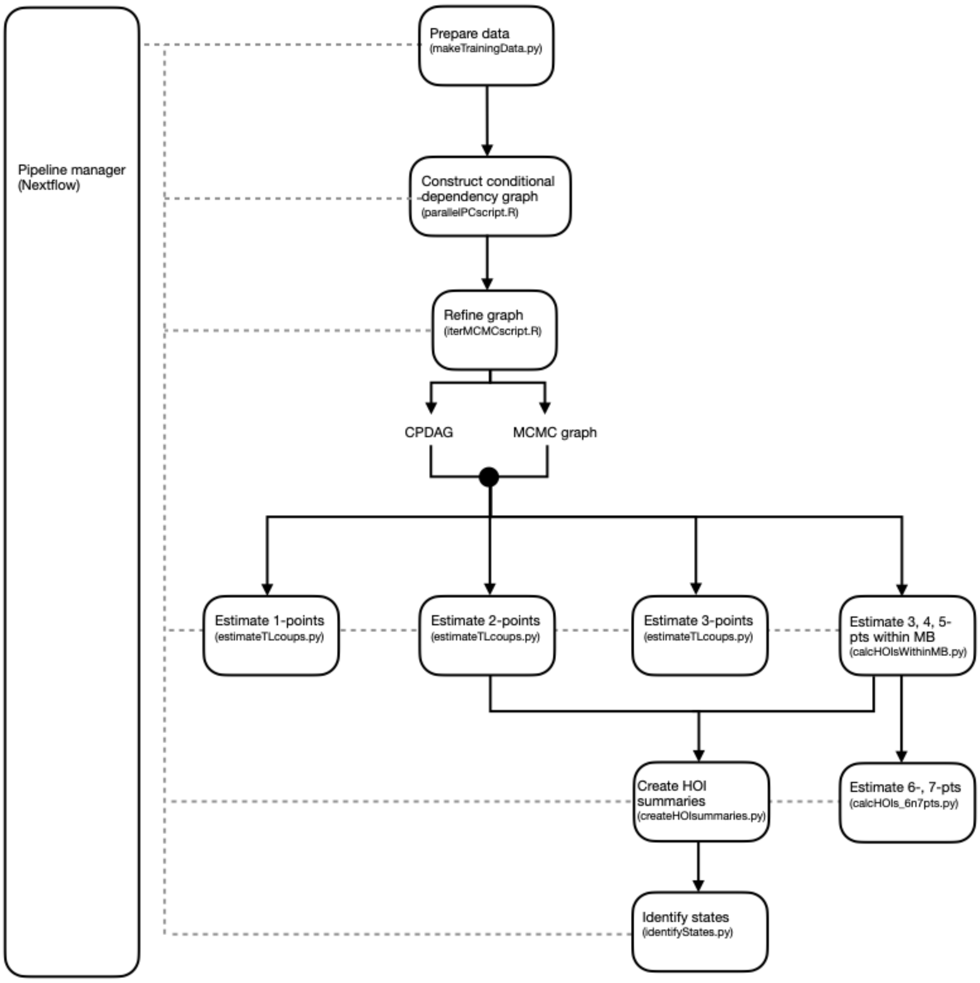


Figure 5.2: The different modules in the `Stator` pipeline.

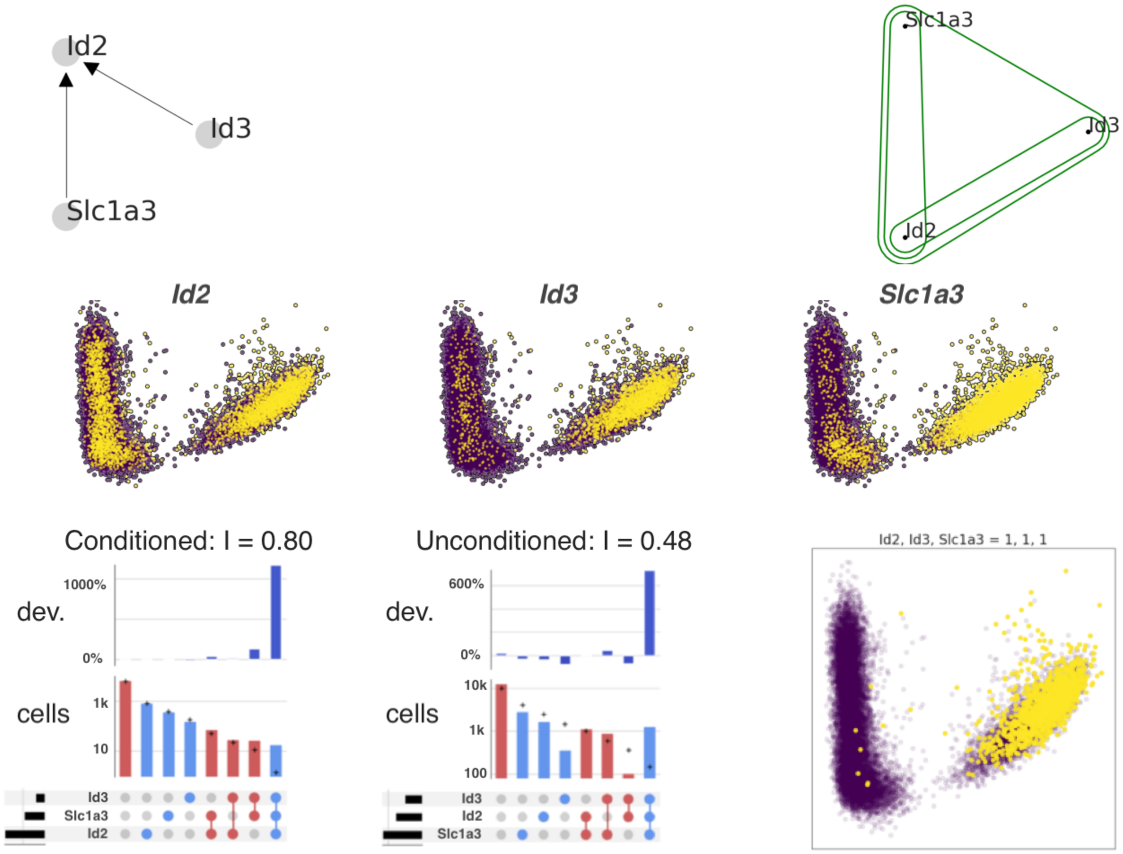


Figure 5.3: An example of a summary of a higher-order interaction, in this case the 3-point interaction among *(Id2, Id3, Slc1a3)*, estimated using the MCMC graph, on a merged data set of developmental neurons and astrocytes. The top left shows the local structure of the MAP CPDAG, the top right the hypergraph of MFIs. The middle row shows the expression of each of the interacting genes in PCA space. The bottom row contains UpSet plots (see Figure 5.21 for more details on such plots) of both the conditioned joint state (bottom left) and the unconditioned joint state (middle). Finally, the bottom right panel shows which cells were in the interaction's characteristic state, which is (1, 1, 1) in this case.

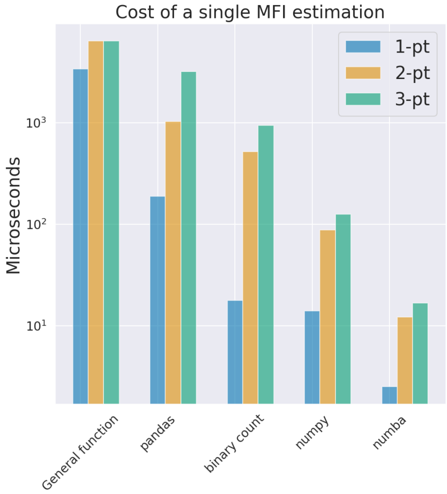


Figure 5.4: Time taken on a single MFI estimation using different estimation methods. Shown for the 1-point of *Id2*, the 2-point of *Id2* and *Etv1*, and the 3-point of *Id2*, *Etv1*, and *Atp1b1*.

```
[8a/e33daf] process > estimatePCgraph (1)               [100%] 1 of 1
[16/5d6bd5] process > iterMCMCscheme (1)                [100%] 1 of 1
[64/d4717f] process > estimateCoups_1pts (2)            [100%] 2 of 2
[ee/3a493f] process > estimateCoups_2pts (2)            [100%] 2 of 2
[ea/70f990] process > estimateCoups_3pts (1)            [100%] 2 of 2
[99/32d387] process > estimateCoups_345pts_WithinMB (1) [100%] 1 of 1
[4e/38e8e2] process > estimateCoups_6n7pts (1)          [100%] 1 of 1
[7d/cd582a] process > createHOIsummaries (1)            [100%] 1 of 1
[da/3899c4] process > identifyStates (1)                [100%] 1 of 1
```

The `eddie_singularity` profile is a custom profile that makes sure the SGE scheduler requests the correct resources, and should be applicable to most SGE schedulers. The `params.json` file contains both the hyperparameters for the algorithms and the parameters used to request resources from the scheduler. An example is shown below.

Code 5.2: Stator parameter file (example)

```
{
"dataType"    : "expression",
"rawDataPath" : "/path/to/data/10X_astrocytes_mouseB.csv",
"nGenes"      : 20,
"nCells"      : 1000,
"fracMito"    : 0.12,
"fracExpressed": 0.02,
"PCalpha"     : 0.05,
"bsResamps"   : 100,
"edgeListAlpha": 0.05,
"nRandomHOIs" : 10,
"minStateDeviation": 5,
"stateDevAlpha": 0.05,
"dendCutoff"  : 0.88,
"auThreshold" : 0.4,
"bsResamps_HC": 100,
"sigHOIthreshold" : 0.05,
"userGenes"   : "/path/to/userGenes.csv",

"executor"    : "sge",
"maxQueueSize"  : 25,
"cores_makeData": 1,
"cores_PC"      : 6,
"cores_MCMC"    : 2,
25  "cores_1pt"     : 4,
26  "cores_2pt"     : 12,
27  "cores_3pt"     : 8,
28  "cores_HOIs_MB" : 6,
29  "cores_HOIs_6n7" : 6,
30  "cores_HOIs_plots": 6,
31
32  "mem_makeData"  : "32G",
33  "mem_PC"        : "16G",
34  "mem_MCMC"      : "16G",
35  "mem_1pt"       : "32G",
36  "mem_2pt"       : "64G",
37  "mem_3pt"       : "64G",
38  "mem_HOIs_MB"   : "64G",
39  "mem_HOIs_6n7"   : "16G",
40  "mem_HOIs_plots" : "64G",
41
42  "time_makeData" : "1h",
43  "time_PC"       : "1h",
44  "time_MCMC"     : "1h",
45  "time_1pt"      : "1h",
46  "time_2pt"      : "1h",
47  "time_3pt"      : "1h",
48  "time_HOIs_MB"  : "1h",
49  "time_HOIs_6n7"  : "1h",
50  "time_HOIs_plots"  : "1h"
51  }
```

The parameters set in line 2−18 correspond to hyperparameters of the various algorithms and can affect the results. The parameters below line 18 all correspond to computational resource requests, so should not affect the results, but might affect the scheduling and execution time. More information on the parameters is available in the documentation at github.com/AJnsm/NF_TL_pipeline/blob/develop/README.md and a vignette with a full walkthrough of the pipeline is available at https://github.com/AJnsm/NF_TL_pipeline/blob/main/vignette/Vignette.md.

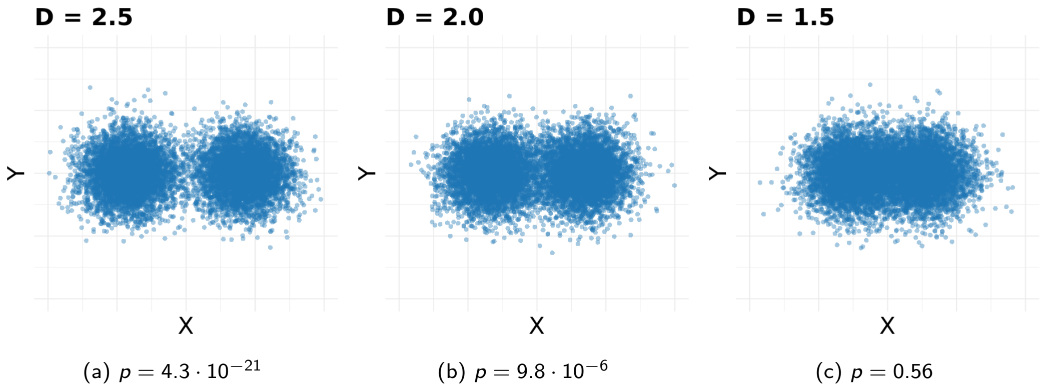
5.2.4 Correcting for double-dipped clusters

When clustering the cells by distance in expression space and then testing for differential (mean) gene expression between these clusters, the same data is used twice in a process known as selective inference, or *double-dipping*. This makes the results of the differential expression analysis unreliable, and it becomes impossible to statistically justify the clustering. In [88], the authors introduce a test for a difference in means that corrects for double-dipping. Given a realisation x of X, where X is a matrix-normally distributed random variable of n observations of q features as $X \sim \mathcal{N}_{n\times q}(\mu, I_n, \sigma^2 I_q)$, one assigns a p-value to a difference in mean across two clusters C_1 and C_2 within a partitioning $C(x)$ as:

$$P_{H_0}\left(||\overline{X}_{C_1} - \overline{X}_{C_2}|| \geq ||\overline{x}_{C_1} - \overline{x}_{C_2}|| \,\middle|\, C_1, C_2 \in C(X) \right) \tag{5.11}$$

That is,

> “Among all realizations of X that result in clusters C_1 and C_2, what proportion have a difference in cluster means at least as large as the difference in cluster means in our observed data set, when in truth $\mu_{C_1} = \mu_{C_2}$?” [88].



(a) $p = 4.3 \cdot 10^{-21}$ (b) $p = 9.8 \cdot 10^{-6}$ (c) $p = 0.56$

Figure 5.5: Data generated from two standard normal distributions (unit variance) a distance D apart is significantly separable for $D = 2.5$ and $D = 2.0$, but not for $D = 1.5$. The p-values are corrected for double-dipping using `clusterpval`. This illustrates that a difference in means of ~ 1.5 is typical for clusters in such data, even when the data is generated from a homogeneous distribution.

To test for a difference in means across clusters, I used the implementation of this test in the `R` package `clusterpval` [88]. I validated the behaviour by generating 5,000 observations from two standard normals ($\sim \mathcal{N}(0, 1)$), shifted to be a distance D apart. I performed a hierarchical clustering (euclidean distance with average linkage) and cut the dendrogram so that there were two clusters of similar size. For small D, the p-value should reflect that there is no statistically sound way to justify separating the population into two. Figure 5.5 shows that this is indeed the case. Note, however, that this is an artificial situation in two dimensions. In the rest of this chapter, I clustered in 20 dimensions, so that a 2D visualisation no longer accurately reflected the dispersion of the data.

5.3 Results

> Things do not exist until they begin to appear.
>
> Commonly attributed to Humberto Maturana

5.3.1 Higher-order characteristic states find structure beyond clustering based approaches

In this section, I compared the resolving power of the characteristic states of higher-order interactions with that of clusters in expression space, correcting the expression space clustering for double-dipping as outlined in Section 5.2.4.

5.3.1.a Interaction-driven clustering

The characteristic states were generated from all Markov-connected 3-, 4-, and 5-point interactions significant at $F \leq 0.05$. I only kept characteristic states that deviated more

Developmental		Adolescent	
Neurons	Astrocytes	Neurons	Astrocytes
62/82 (76%)	76/98 (78%)	45/50 (90%)	9/10 (90%)

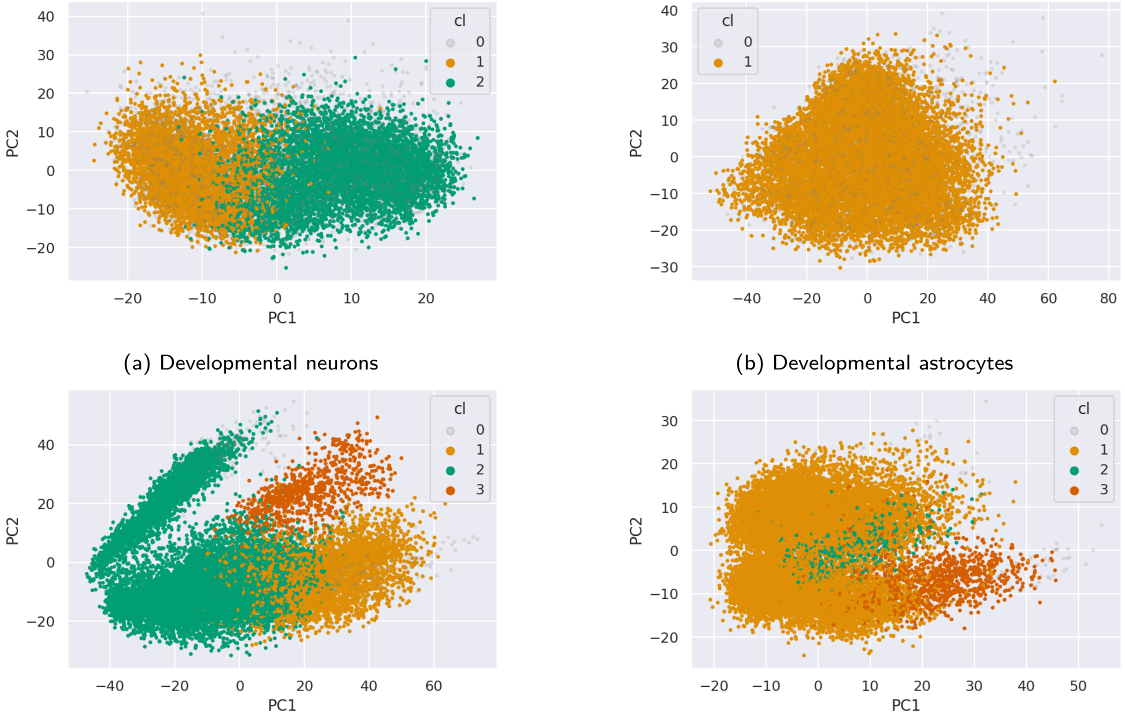
Table 5.1: Shown is the fraction of characteristic states that are $\widehat{1}$ for the developmental and adolescent data sets.

than a factor of 5 at a significance level beyond $p \leq 0.05$ with respect to the Bernoulli null. In practice, for the cells in these data sets, all states that deviated more than a factor of 5 were significant at that level. The characteristic states tended to be the states where all interacting genes were expressed, denoted $\widehat{1}$. Across the four data sets, I found between 10 and 98 characteristic states, most of which corresponded to the $\widehat{1}$ state. A summary of the states is shown in Table 5.1.

5.3.1.b Data-driven clustering

Cells are often clustered in expression space with the Louvain clustering algorithm, but that optimises for modularity, not total cluster number. As I am primarily interested in maximising resolving power, I created a standard hierarchical clustering in the first 20 principal components (PCs) of the expression data. The elbow of the explained variance by the PCs occurs well below 20 PCs in each of the data sets (Figure 5.7). To cut the dendrogram at a height that maximises the total number of significant clusters, I ran a series of tests on subsampled data across a range of cutting heights. Using the first 20 PCs, a random selection of 5,000 cells from the data set was clustered hierarchically (using a Euclidean distance and average linkage). In each data set I generated 5 clusterings by cutting the dendrogram at $k = 2$, 4, 6, 8, or 10, where k is the total number of clusters. For each of these clusterings, I calculated the significance of the difference in means using the `clusterpval` package (see Section 5.2.4), and only kept clusters that contained at least 2.5% of all cells, and had a significant (at $\alpha = 0.05\frac{k(k-1)}{2}$) difference with at least one other cluster. I then chose the final number of clusters by fixing k to the lowest value that had as many significant clusters as the $k = 10$ clustering. This way, the number of meaningful, *i.e.* significant, clusters was maximised, while minimising the number of cells that cannot be assigned to a significant cluster. For both developmental data sets, this corresponded to setting $k = 6$, for the adolescent neurons to $k = 8$, and the adolescent astrocytes never yielded more than 1 significant clustering. I chose not to increase beyond $k = 10$ because the significance calculations are unreliable when the clusters become too small. Using these values for k (and setting $k = 8$ for the adolescent astrocytes to match the neurons), I clustered all the cells in each data set, and called this clustering the *data-driven* clustering. This resulted in respectively (2, 1, 3, 3) clusters for the developmental neurons and astrocytes, and adolescent neurons and astrocytes, shown in Figure 5.6.

The data-driven approach yielded fewer than three clusters in each data set, while the interaction-driven approach yielded an order of magnitude more states. Note, however, that it would be a simplification to say that the interaction-driven approach offered a higher resolution description, since it is still unclear if the interaction-driven states are a fine-grained description of the cell types usually found by expression space clustering



(a) Developmental neurons

(b) Developmental astrocytes

(c) Adolescent neurons

(d) Adolescent astrocytes

Figure 5.6: Data-driven clustering of the four data sets, embedded in the first two principal components. If a cell was assigned to a cluster that contains fewer than 2.5% of cells, or if there was no significant difference in mean with any other cluster, then it got assigned to cluster 0, here shaded in grey.

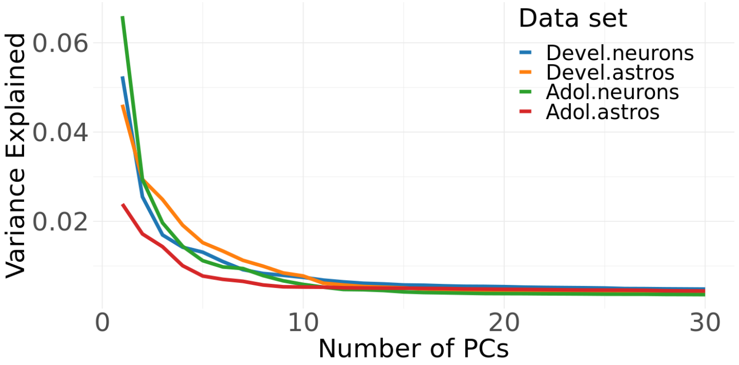
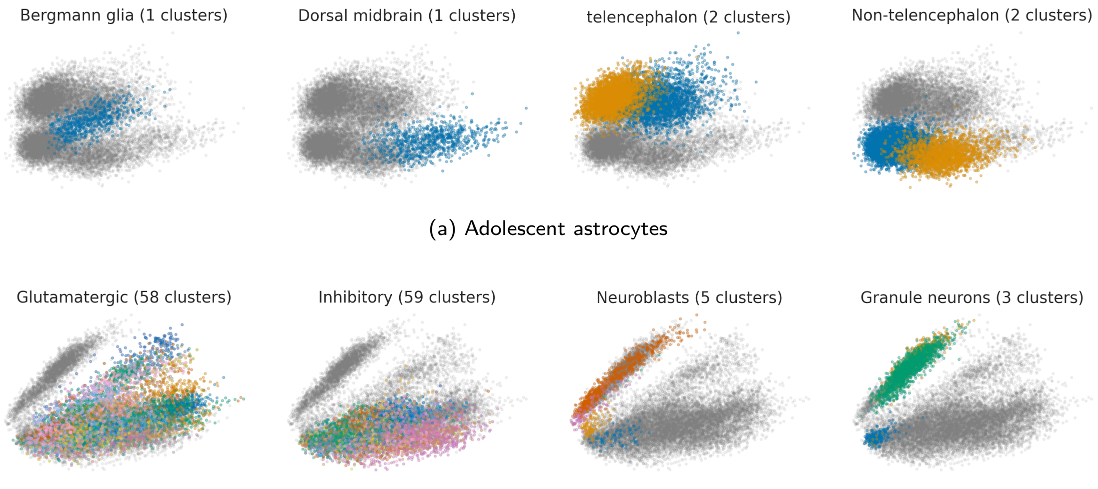


Figure 5.7: In all four data sets, the first 10 to 15 principal components (PCs) explain the most variance, so that including 20 PCs in the clustering should capture most structure.



(a) Adolescent astrocytes

(b) Adolescent Neurons

Figure 5.8: Expert-driven clustering of the two adolescent data sets. Each colour corresponds to a distinct expert annotation, so the actual clustering is finer than the eight clusters shown in this figure, which just serves to give a broad picture of the embedding of each subtype.

algorithms, or a new kind of biological state altogether. This will be explored in the next section by comparing the interaction-driven states with expert annotations.

5.3.2 Combinatorial markers for known cell types

5.3.2.a Expert-driven clustering

The adolescent data used in this thesis was part of the Mouse Brain Atlas [313], and thus came with cell type annotations. This annotation will be referred to as the *expert annotation* throughout this thesis. These annotations were based on a combination of hierarchical clustering, Louvain clustering, k-nearest neighbour clustering, density-based algorithms, trained classifiers, and the manual merging and elimination of clusters by experts. According to this clustering, the neurons had cells from 157 subtypes, and the astrocytes contained 7 subtypes. Each cell in the two adolescent data sets was annotated with this subtype, which I called the *expert-driven* clustering, and a rough overview is given in Figure 5.8. Inspection by eye showed that the data-driven clusters were all also found by the expert-driven approach, but the expert-driven approach had a higher resolution. It was not immediately clear how to interpret the expert-driven approach, as it finds more structure than can be justified by mean gene expression, while still being based on mean expression in the sense that the clusters necessarily localise in expression space. It should, however, be noted that the authors of [313] used a more sophisticated dimensional reduction technique than taking the first few principal components. Still, the discrepancy between the 3 significant clusters after correcting for double-dipping and the 157 clusters found in the expert-driven approach is striking.

5.3.2.b Interaction-driven states can be cell-type specific

To see if the characteristic states corresponded to the fine-grained description from the *expert-driven* clustering, I calculated the overlap in annotation between the two methods. Let the set of cells in a characteristic state S have an overlap of size S_i with cluster C_i, such that its specificity σ_i and sensitivity τ_i with respect to cluster C_i can be defined as

$$\sigma_i = \frac{S_i}{\sum_j S_j} \tag{5.12}$$

$$\tau_i = \frac{S_i}{\mid C_i \mid} \tag{5.13}$$

Of particular interest were the clusters that the characteristic states are most specific and sensitive to, so define:

$$\sigma = \max_i \sigma_i \tag{5.14}$$

$$\tau = \tau_{\text{argmax } \sigma_i} \tag{5.15}$$

I then calculated τ and σ of all characteristic states in the adolescent data (since the expert-driven clustering was only available for the adolescent cells).

Adolescent astrocytes Keeping only those characteristic states in the adolescent astrocytes with a Youden index (sensitivity + specificity −1) above 0.05 for any of the expert annotated clusters, I found the three characteristic states in Table 5.2. Each of the three states was more than 90% specific to the ACBG class, which corresponded to the Bergmann glia of the cerebellum—indeed a subtype of astrocytes. In the Mouse Brain Atlas, *Gdf10* was already annotated as a marker for ACBG, but demanding only *Gdf10* to be expressed resulted in a specificity of just 0.42 and a sensitivity of 0.82. Two of these three states were derived from 3-point interactions with *Hopx* and *Gria4*, and one of two calcium-dependent genes: *Cpne2* or *Camk1*. *Gria4* encodes the protein GluR4, which is part of the transmembrane AMPA receptor that mediates calcium permeability. The CPDAG and the MCMC graph are identical in both cases, and locally form colliders on the calcium genes as *Gria4*→*Cpne2*←*Hopx* and *Gria4*→*Camk1*←*Hopx*. This is consistent with the two calcium-dependent genes being regulated by the transcription factor *Hopx*, and functioning conditional on the presence of GluR4. However, this interpretation is contingent on the quasi-causal graph reflecting the true causal structure in both these cases, so should be investigated further by validating the presence of this dependency in multiple other data sets before any definite conclusions can be reached. Note that both these states come with a high specificity, but low sensitivity. This suggests that they might reveal a particular substate of the Bergmann glia where the calcium is available. Moreover, the set of cells in the two states are not equal—they have an overlap coefficient ($A \cap B / \min(\mid A \mid, \mid B \mid)$) of 0.51, indicating that they define two partially disjoint sets of cells. The third characteristic state results from a 4-point interaction with the Bergmann glia markers *Gdf10* and *Sept4*. The third gene present is *Tlcd1* (Calfacilitin), another calcium channel modulator. Interestingly, the fourth gene in the interaction, *Igfbp2*, is absent in the characteristic state. *Igfbp2* binds to the calcium regulating IGF

genes	state	Expert annot.	σ	τ
Tlcd1, Sept4, Gdf10, Igfbp2	1, 1, 1, 0	ACBG	0.92	0.58
Cpne2, Gria4, Hopx	1, 1, 1	ACBG	0.92	0.31
Camk1, Gria4, Hopx	1, 1, 1	ACBG	0.90	0.16

Table 5.2: Characteristic states in the Zeisel astrocytes.

proteins, but has also been shown to affect intracellular calcium concentrations in a human cancer cell line (MCF-7 cells) [236]. All three Bergmann glia-specific states thus revolve around calcium channel regulation. That regulation of calcium permeability is crucial to neurogenesis and the development and functioning of Bergmann glia has been previously established in mice [166, 113].

I next undertook the same analysis using the data-driven clustering of the adolescent cells. Keeping only those characteristic states with a Youden index above 0.05 for any of the data-driven clusters, I found the same three characteristic states in Table 5.2 that are all at least 90% specific to cluster 2. Cluster 2 indeed corresponds to the cluster of Bergmann glia in the expert annotation, as can be seen by comparing Figures 5.8 and 5.6, and by the fact that the ACBG expert annotation and the data-driven cluster 2 had an overlap coefficient of 0.99. Data-driven clusterings thus provided no more information than the expert-driven clustering.

Adolescent neurons The characteristic states in the adolescent neurons with Youden index above 0.05 with respect to the expert annotation are shown in Table 5.3. Most of these were specific to the various clusters of glutamatergic neurons of the spinal cord (those with expert annotation SCGLU[X]), and among the genes that appeared in these interactions (*Penk, Tac1, Sst, Grp, Pthlh, Gad2, Hoxb8*), only *Gad2* and *Hoxb8* were not neuropeptide genes (though *Gad2* only appears as a negative marker). Almost all these neuropeptide gene interactions were specific to the cluster SCGLU4, and all involved *Penk* and *Hoxb8*. One interaction was specific to SCGLU5 and indeed involved different neuropeptides. All but one of the SCGLU4 specific interactions involved *Sst*, which was surprising considering the fact that mean expression of *Sst* is used to mark GABAergic neurons in the cortex (see [250] and `mousebrain.org/adolescent/genes.html`[313]), though other research has found the corresponding mRNA molecules to be expressed across all types of neurons [248].

The interactions specific to neuroblasts (expert annotation DGNBL2: granule neuroblasts, dentate gyrus) involved *Sox11* and *Igfbpl1*, both canonical markers for that cell type. The $\widehat{1}$-state of the genes *(Id2, Tcf4, Igfbpl1)* was particularly interesting as it involved both a canonical neuroblast marker and the pair *(Id2, Tcf4)* which the String Database annotates as interacting with the highest level of confidence (> 0.90), in part based on 'Experimental/Biochemical data' [264]. Finally, one interaction was specific to inhibitory neurons in the spinal cord (SCINH6), and also involved (neuro)peptides, and the calcium-signalling gene *Calb2* whose protein is known to interact with neuropeptides and has been used in combination with *Npy* to mark GABAergic neuronal subtypes in mice before [153].

genes	state	Expert annot.	σ	τ
Penk, Tac1, Sst, Hoxb8	1, 1, 1, 1	SCGLU4	0.82	0.78
Grp, Pthlh, Hoxb8	1, 1, 1	SCGLU5	0.94	0.70
Penk, Tac1, Sst	1, 1, 1	SCGLU4	0.56	0.80
Penk, Sst, Hoxb8	1, 1, 1	SCGLU4	0.51	0.80
Penk, Tac1, Hoxb8	1, 1, 1	SCGLU4	0.74	0.93
Nppc, Npy, Calb2	1, 1, 1	SCINH6	0.57	0.87
Id2, Tcf4, Igfbpl1	1, 1, 1	DGNBL2	0.66	0.41
Sox11, Tubb2b, Igfbpl1	1, 1, 1	DGNBL2	0.71	0.80
Penk, Sst, Gad2, Hoxb8	1, 1, 0, 1	SCGLU4	0.53	0.80
Tnnt1, Rora, Lhx1os	1, 1, 1	MEINH5	0.48	0.71
Hist1h2bc, Zbtb20, Ddn	1, 1, 1	DGGRC2	0.82	0.30

Table 5.3: Characteristic states in the adolescent neurons.

In conclusion, the characteristic states seemed to reproduce some of the expert-driven cell types, but not precisely. In the adolescent astrocytes, the characteristic states were highly specific ($\sigma > 0.9$) to the expert annotation of Bergmann glia, but not very sensitive. This is consistent with the characteristic states corresponding to various states or subtypes of the Bergmann glia. This is further supported by the fact that the different states are partially disjoint. In the adolescent neurons, the characteristic states were often moderately sensitive, but less specific. This is consistent with the characteristic states representing states that are present throughout different cell types. An advantage of the characteristic states is that they are associated to a limited set of genes that can help with the interpretation of the states. For example, the states in the Bergmann glia all involved regulation of calcium transport, while those in the adolescent neurons were almost always defined by their neuropeptides. Regardless of their biological interpretation, that the states are more specific in the astrocytes and more sensitive in the neurons is to be expected considering the difference in homogeneity between the two cell types. The expert annotation clusters in the neurons have a median of 57 cells per cluster (IQR=18 – 116 cells), whereas the astrocyte annotations had a median of 2,148 cells (IQR=1124 – 3260 cells).

So far, I have considered each characteristic state as a true state in itself. Since some characteristic states might describe the same biological state, the characteristic states might contain redundant structure, and need to be clustered to reflect coherent and distinct biological states. For example, the *(Penk, Sst, Hoxb8)* characteristic state seemed to correspond to a similar set of cells as the *(Penk, Sst, Gad2, Hoxb8)* characteristic state, since they were both very similarly specific and sensitive to the SCGLU4 cluster. In fact, adding the *Gad2* condition could only decrease the total number of cells in that state, so the cells in the *(Penk, Sst, Hoxb8)* state were a superset of the cells in the *(Penk, Sst, Gad2, Hoxb8)* state.

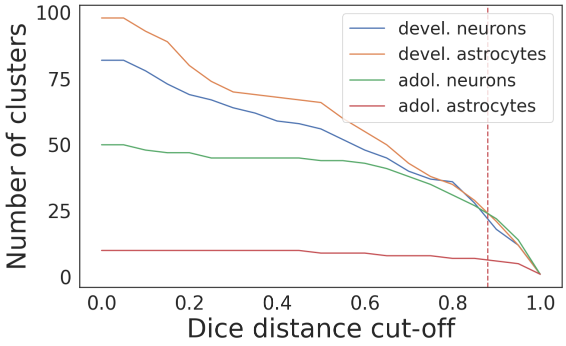


Figure 5.9: The mean Dice distance at which the elbow occurred in the cut-off response curve was 0.88, here shown in red (dashed). An added benefit of this is that this cut-off results in a similar number of clusters in three of the four data sets.

5.3.3 A robust clustering of characteristic states

As outlined in Section 5.2.1, the characteristic states were hierarchically clustered using average linkage and a distance defined by the Dice distance. This was done in each data set separately, and cutting the dendrogram yielded a clustering of characteristic states for each data set. The value of this cut-off was set to 0.88 and kept the same throughout the data sets so that the total number of states could be compared. This threshold was chosen by finding the elbow in the graph that relates the total number of found clusters to the cut-off distance. The elbows occurred at a Dice distance of 0.80 in the adolescent data sets, and at 0.95 in the developmental data sets (judged by eye, summarised in Figure 5.9), leading to a mean distance cut-off of 0.88. In two further `Stator` runs on the same data sets, this mean elbow occurred at 0.84 and 0.81, but I kept the threshold at 0.88 for reproducibility, and because a higher threshold results in fewer states and is thus more conservative. This is, however, an arbitrary and imprecise way to set the cut-off, so Section 5.3.7 outlines a method that constructs the states automatically by quantifying how robust each individual cluster is, abolishing the need for a cut-off to be specified. Another method to automate the choice of cut-off is to maximise the modularity of the final clustering, which is currently the default in `Stator`, and leads to very similar results. In this section, I simply cut the dendrogram and verified that the resulting clustering was robust.

To verify that the resulting clusters were robust with respect to finite sample variability, I bootstrap resampled the cells of the whole data set, and reclustered the characteristic states. Note that I did not recalculate the interactions, as that would conflate the robustness of the interaction estimation with that of the clustering procedure. I resampled the data 1,000 times, and for each resample calculated the adjusted Rand index (ARI) and the adjusted mutual information (AMI), relative to the original clustering (see Section 5.2.2). The results are reported in Figure 5.10 and ranged from good (> 0.8) to excellent (> 0.9). From this, I concluded that the clustering was robust. Therefore, I refer to a cluster of characteristic states simply as a *state*.

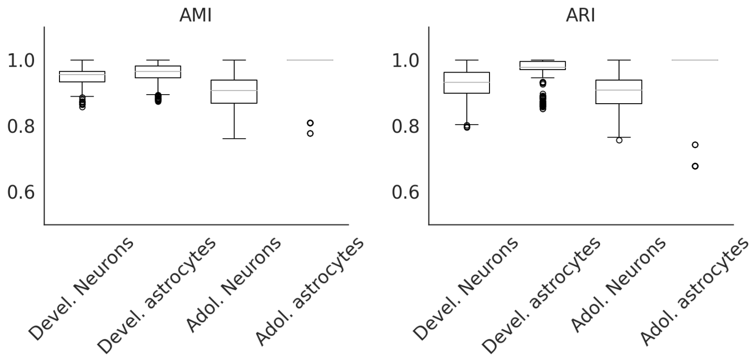


Figure 5.10: The AMI and ARI across 1,000 bootstrap resampled clusterings, relative to the original clustering, ranged from good (> 0.8) to excellent (> 0.9), indicating that cutting the dendrogram at a height of 0.88 led to clusters that were robust with respect to the finite sample variance. The adolescent astrocytes were almost always perfectly robust (AMI=ARI=1.0), but that could be in part attributed to there being many fewer clusters.

5.3.4 Higher-order states reflect diverse biology

In Figures 5.13 to 5.15 the resulting truncated dendrograms are shown, where each leaf is annotated with the four genes that appeared most often in interactions in that cluster (where +/- denotes the presence/absence of that gene, and equally often occurring genes are listed alphabetically), and the embedding of the characteristic states in a PCA embedding. The developmental neurons and astrocytes resulted in a similar number of clustered states—15 and 16, respectively—while the adolescent neurons and astrocytes yielded 23 and 6 clusters, respectively. That the adolescent astrocytes form a much more homogeneous group of cells than the adolescent neurons was already implied by the expert annotation of expression space in cells, and is corroborated here by the clustering of characteristic states. Clustering the developmental neurons and astrocytes by expression resulted in respectively just 2 and 1 significant clusters, but in many more clustered states. To end up with just 2 states in these data sets, the dendrogram would have to be cut at a distance beyond 0.99. Furthermore, it was also immediately clear that these states revealed different structure from expression space clusters, as multiple states did not localise in the first two principal components. As the states correspond—ideally—to biologically meaningful cell identities, such delocalised states correspond to a single identity that is instantiated in different cell types. Therefore, I refer to states that delocalise in expression space as *polysemic* states. To see what each of the states corresponded to, various interesting clusters in the data sets are discussed below in more detail.

Developmental neurons The data driven clustering separated this data set into two clusters along the horizontal principal component (PC1 in Figure 5.6). Figure 5.12 shows that most states also separated along this axis, localising to the right/east of the embedding plane, or the left/west. There were two polysemic exceptions to this: an *Hba/b+* state with *Hbb-bs*, *Hbb-bt*, *Hba-a1* and *Hba-a2* expressed, and a state with *Fos*, *Junb*, and *Nr4a1*. Neurons with hemoglobin gene expression have previously been found in mice and humans [28], where the presence of *Hba/b* transcripts was linked to mitochondrial activity. However, this state could also correspond to multiplet transcriptomes that contain blood cells. Genes from the *Jun* and *Fos* families encode the heterodimer AP-1, whose transcription is induced by neuropeptides and electrical excitation [239], and neuronal plasticity [174]. AP-1 is generally associated to neural activity and might be involved in regulating the cell type specificity of activity-induced gene expression [310]. The third gene, *Nr4a1*, directly regulates *Junb* and *Fos* by binding to their 3' UTR [100], and is also activated in response to neuronal activity [310]. The *Junb/Fos/Nr4a1* state thus seemed to correspond to active neurons.

The states that localised in the east of the PCA embedding separated roughly into *Ebf1/Isl1* positive cells, and *Gucy1[x]3/Six3* positive cells, both of which also expressed *Foxp1*, a marker for medium spiny neurons (MSNs) [198]. Two subtypes of MSNs are commonly described, D1 and D2 MSNs. Both *Ebf1* and *Isl1* are markers for D1 MSNs, but *Six3* and *Gucy1a3* are mostly specific to D2 MSNs [249, 319]. This suggests that the *Ebf1/Isl1*+ state corresponded to D1 MSNs, while the *Gucy1[x]3/Six3*+ state corresponded to D2 MSNs. Furthermore, *Zfp503*, which defined the state most closely related (in terms of Dice distance) to the D1 MSNs, is a necessary transcription factor for D1 MSN differentiation [319].

The states that localised to the west of the PCA embedding roughly separated into two categories: one characterised by the expression of the transcription factor *Etv1*+, and one by *Cenpa/Ube2c/2610318N02RIK/Arhgap11a*+ and *Vim/Ezr/Mdk/Dbi*+ cells.

The genes *Etv1, Arl4d* and *Smoc1* are markers for GABAergic neurons in different parts of the brain, according to the developmental Mouse Brain Atlas [313]. In the same atlas, the genes *Cenpa, Ube2c* and *Arhgap11a* all mark neuronal intermediate progenitor cells (nIPCs), most likely because they mark specific points along the cell-cycle. Furthermore, *Vim* and *Dbi* specifically stimulate proliferation of neuronal progenitor cells [52, 9], and *Mdk* is likewise involved in cell-cycle control [301].

Recall that the data-driven clustering partitioned the cells along PC1 in Figure 5.6 into just two clusters, separating MSNs from the rest of the population. The characteristic states revealed structure beyond this binary classification. The MSNs separated into D1 and D2 medium spiny neurons, while the states in the west of the embedding separated into a different class of GABAergic neurons, and a cycling, progenitor state. Crucially, the characteristic states did not just reveal more structure, it revealed delocalised cell identities that are *fundamentally* impossible to detect by clustering in expression space.

Developmental astrocytes The data-driven approach yielded no significant clusters in this data (see Figure 5.6), so clustering in expression space gave no statistically sound way to conclude that there is any substructure present in this cell population. In contrast, the characteristic states of the higher-order interactions revealed 15 cell states, shown in Figure 5.13. Farthest removed from all others, there was an *Hba/b*+ state present throughout expression space, similar to the one found in the neuronal population, but here involving *Foxj1* (a gene that has been shown to regulate other kinds of Globin proteins [138]). The authors of [28] also found astrocytes that expressed these hemoglobin genes, but again, this state could also correspond to blood cell containing multiplets.

Beyond that, a range of *Cenpa*+ states clustered together. The set of genes that appeared in the *Cenpa/Pbk/Cdca3/Lockd*+ state was more than 12-fold enriched in the GO:BP terms *organelle fission, nuclear division, chromosome segregation, cell division* and *nuclear chromosome segregation* (all significant with a g:SCS corrected p-value < 0.022, using the 1k HVGs as a background), and thus indicated a mitotic cell state around the G2M transition. Note that *Lockd*, which also appeared in this state, is a lncRNA that itself has no gene ontology annotation, but is directly downstream of *Cdkn1b* [184], another cell-cycle regulation gene (in fact, the official name for *Lockd* is *lncRNA downstream of Cdkn1b*). The *Gins2/Fen1/Ung/Rpa2*+ state separated from the *Cenpa*+ state at a high Dice distance and was more than 42-fold enriched in the GO:BP terms *DNA Replication, Base excision repair, cell cycle DNA replication, DNA unwinding involved in DNA replication, nuclear DNA replication, and base-excision repair* (all significant with a g:SCS corrected p-value < 0.017, using the 1k HVGs as a background). This state thus seemed to correspond to cells in S-phase. The interactions in the S-phase state involved the genes *2810417H13Rik/Pclaf*, ***Dhfr***, ***Fen1***, ***Gins2***, *Gmnn*, *Hells*, *Lig1*, ***Mcm5***, ***Mcm6***, *Pcna*, ***Plk4***, ***Rad51***, ***Rpa2***, *Rrm2*, ***Smc2***, ***Top2a***, *Uhrf1*, ***Ung*** and those in the G2M states (the three rightmost states in the dendrogram) involved *2810417H13Rik/Pclaf*, ***Ccnb2***, ***Cdca3***, ***Cdca8***, ***Cdkn3***, ***Cenpa***, ***Cenpf***, *Cenph*, *Gmnn*, ***Knstrn***, *Lockd*, *Pbk*, ***Racgap1***, *Rrm2*, ***Tacc3***, ***Top2a***, ***Ube2c***, *Uhrf1*. All genes printed in bold have been identified as targets of the p53-DREAM pathway in [76], which regulates gene expression during the cell-cycle. The DREAM complex binds to promoters with E2F binding sites, which were already identified in Chapter 4 as enriched in the set of interacting genes across the four data sets. Moreover, the p53-DREAM target genes preferentially coupled together in MFIs. That is, the genes that are not known to be p53-DREAM targets coupled mostly together in the interactions *Cenph-2810417H13Rik-**Cenpa***, *Gmnn-Pcna-Lig1*, *Gmnn-Rrm2-**Cenpa***, *Hells-Lig1-**Gins2***, *Hells-Rrm2-**Gins2***, and *Cenph-Uhrf1-**Cenpa***. This, and the fact that higher-order interactions were enriched in the E2F binding sites, suggested that the differences between the cell-cycle states were in part driven by differential p53-DREAM regulation.

Separated from these cell-cycle related states, two closely related states appeared, involving *Neurod2/Neurod6/Auts2/Dcx*+ and *Bcl11b/Dcx/Meg3/Stmn2*+ cell. All of these genes are canonically involved in neuronal differentiation or maturation, except for *Meg3* (Maternally expressed gene 3), a maternally expressed lncRNA. However, there is a possible explanation for its involvement, specifically in neuronal differentiation among a population of astrocytes. *Meg3* is flanked on the chromosome by the protein-coding gene *Dlk1* (and the lncRNA *Meg8/Rian*). *Dlk1* is known to be important to neurogenesis [258], but it is usually only paternally expressed, whereas *Meg3* is maternally expressed, which makes it unlikely that they are coexpressed simply due to their genomic proximity. However, it has been shown that *Dlk1* is essential to neurogenesis in a maternally expressed form in niche astrocytes and neural stem cells [80]. There is no *Dlk1*-associated state, but that could be because *Dlk1* was only lowly expressed (in 2% of cells, *vs.* 18% of cells with *Meg3* expression), had no estimable 3-point interactions, and only two perfectly significant estimable 2-point interactions (with *Miat* and *Fbln2*). However, there was a weak but significant Pearson correlation between *Meg3* and *Dlk1* ($r = 0.18$, $p = 6 \cdot 10^{-120}$). Therefore, the *Meg3* expression in this neuronal differ-

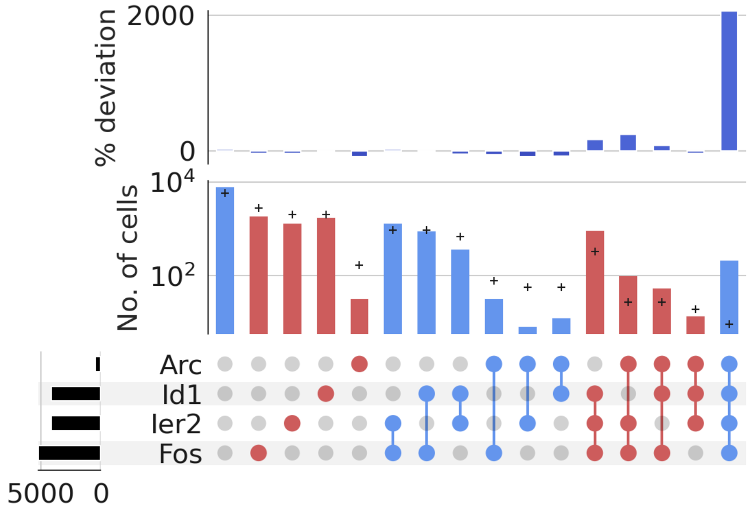


Figure 5.11: UpSet plot of the *(Arc, Id1, Ier2, Fos)* interaction. Plusses indicate the expected number of cells under the Bernoulli null. The $(1, 1, 1, 1)$-state was more than 2,000% overrepresented relative to the Bernoulli null. It can also be seen that *Arc*-transcripts were found more often in combination with all three other transcripts than by themselves.

entiation state might reflect the role of maternally expressed *Dlk1* in niche astrocytes. Bringing these observations together, these two states might correspond to a neurogenesis state in radial glial cells in a microenvironment of niche astrocytes, explaining why there was a neurogenesis signature in a cluster that by differential expression only showed astrocyte markers.

The state defined by the 3 genes from the *Id* family and *Sparcl1* was already observed in [151], where they were identified as astrocyte progenitor cells (specifically referred to as APC2). In the dendrogram, it was closely related to a state defined by two interactions: *Arc-Id1-Ier2-Fos* and *Fos-Dusp1-Ppp1r15a*. Both these states involved only immediate-early genes (except for *Ppp1r15a*), and in particular *Fos* which is known to mark a subtype of astrocytes known as *immediate-early* astrocytes (ieAstrocytes) [99]. *Arc* is usually expressed in neurons rather than astrocytes, but *Arc* codes for a protein that can form virus-like capsids responsible for the intracellular transport of mRNA [186]. In particular, neuronal activity stimulates the transport of *Arc*-transcripts from neurons into astrocytes [186]. Arc was indeed only lowly expressed in the astrocytes, but the *Arc/Id1/Ier2/Fos+* state was more than 2,000% enriched relative to the Bernoulli null (see Figure 5.11 for a summary of the different expression patterns). There were two distinct biological explanations for this observation. First, it could be that *Id1, Ier2* and *Fos* transcripts were transported along with *Arc* mRNA inside the Arc-vesicles. The authors of [186] indeed found that—in a bacterial model—Arc-vesicles could transport other transcripts as well. Alternatively, the conditional expression pattern could have been the result of endogenous transcription of IEGs in astrocytes in response to the presence of *Arc* containing vesicles.

Adolescent neurons To interpret the characteristic states in the adolescent data sets, I compared them with the expert annotations. By representing each expert-driven cluster

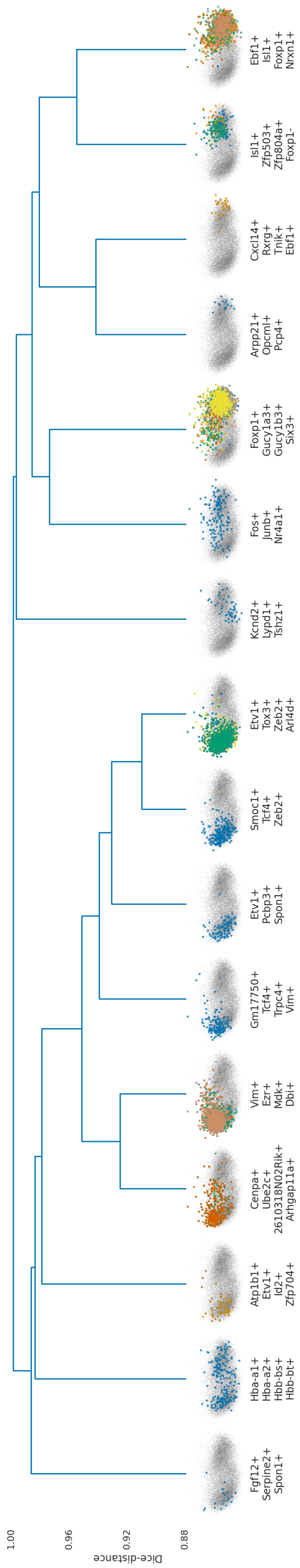


Figure 5.12: Hierarchical clustering of developmental neuron states.

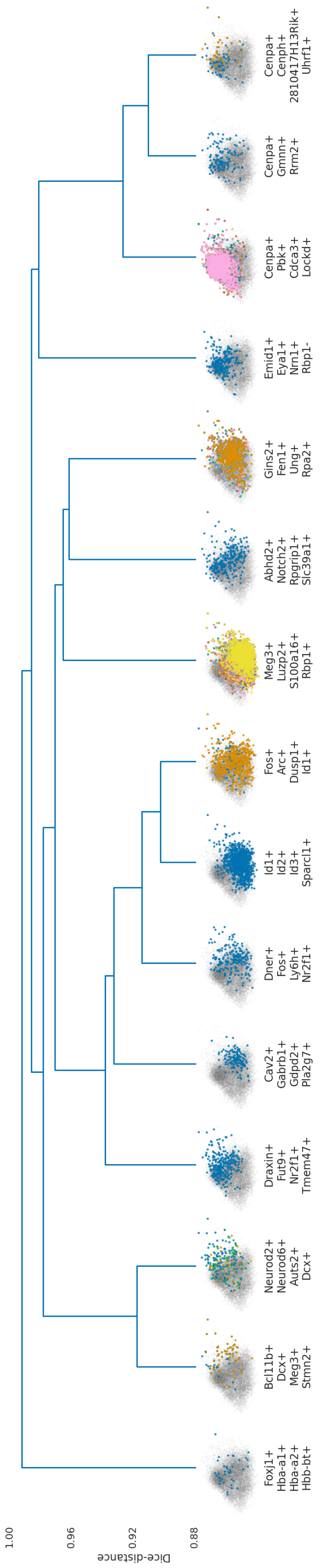


Figure 5.13: Hierarchical clustering of developmental astrocyte states.

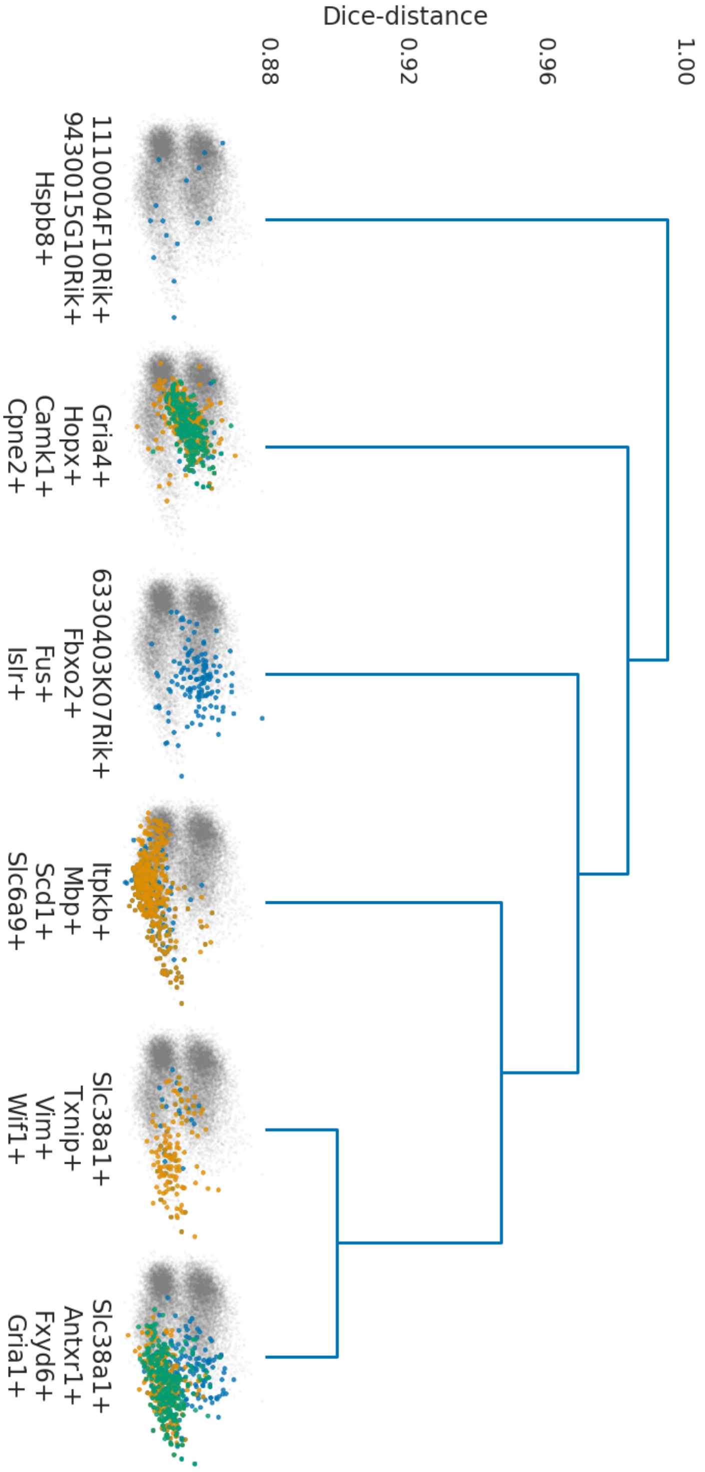


Figure 5.14: Hierarchical clustering of adolescent astrocyte states.

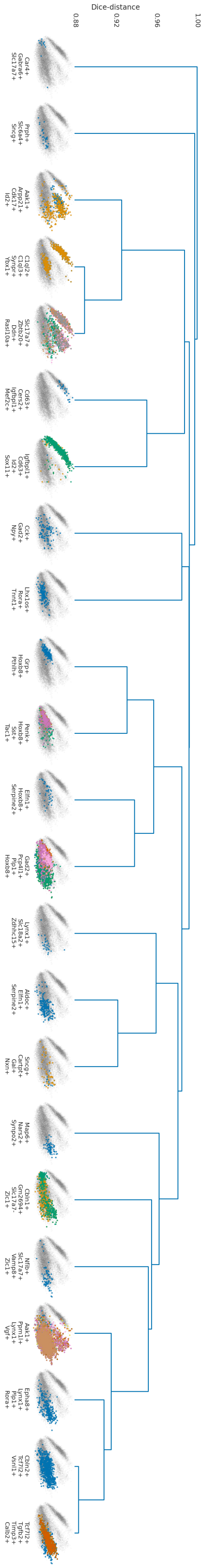
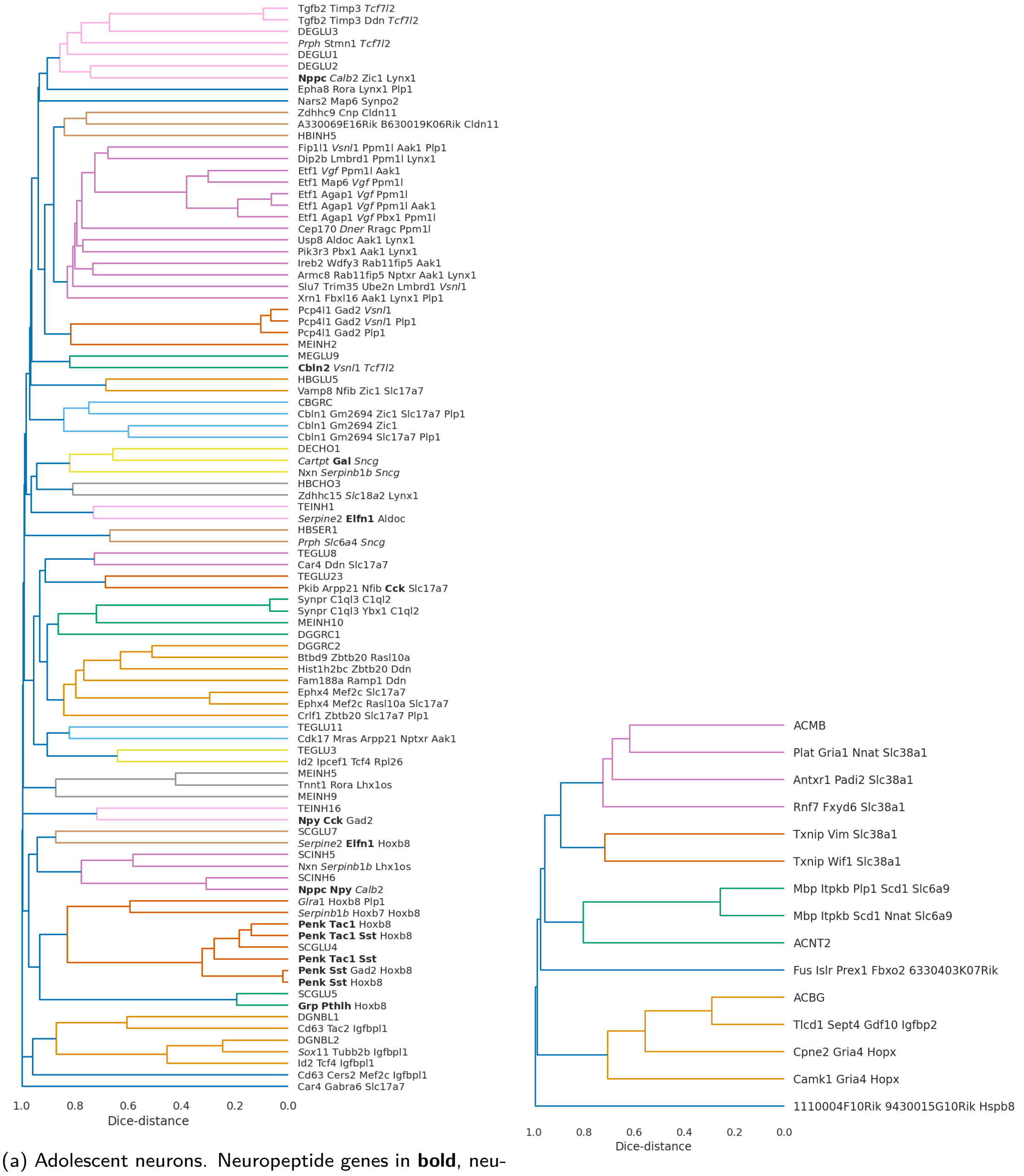


Figure 5.15: Hierarchical clustering of adolescent neurons states.



(a) Adolescent neurons. Neuropeptide genes in **bold**, neuropeptide signalling genes in *italics*.

(b) Adolescent astrocytes

Figure 5.16: A hierarchical clustering of the higher-order interactions and the expert annotation. Leaves that were less than a Dice distance of 0.88 removed were grouped by colour. Only expert annotations that did not end up as a singleton cluster after cutting the dendrogram at a Dice distance of 0.88 are shown.

in the same way as the characteristic states (*i.e.* a binary vector that indicates if a particular cell has that annotation), I could recluster the characteristic states while including the expert annotations as additional states. This revealed how characteristic states overlapped with the expert-driven annotations. The result of this new clustering is shown in Figure 5.16. In theory, this reclustering could have changed which interactions ended up in the same cluster, but in practice proved quite robust (AMI=0.83, ARI=0.78). Of the 13 singleton states in the clustering of Figure 5.15, only 4 remained singletons in the new clustering of Figure 5.16, which shows that most singleton states overlapped significantly with a particular expert annotation. The leaves of the dendrogram are labelled by the genes involved, and a gene name is printed in bold when it encodes a neuropeptide, and in italics when it is a neuropeptide signalling molecule as defined by orthology with cluster 73: *Neuropeptide signalling* from the tissue expression map in the Human Protein Atlas (`https://www.proteinatlas.org/humanproteome/tissue`) [278].

There were 9 clusters in Figure 5.16 strongly enriched in neuropeptides (shown in bold), and peptidase inhibitor genes (*Serpine2* and *Serpinb1b*), and each associated to an expert annotation for glutamatergic, GABAergic/inhibitory, cholinergic, or serotonergic neurons. When a cluster contained multiple expert annotations (for example the cluster associating with DEGLU1, DEGLU2, and DEGLU3), they were always of neurons with identical neurotransmitter identities, consistent with neuropeptides and neurotransmitters being coreleased preferentially in certain combinations. While *Gad2*, which encodes a GABA-synthesising protein, appeared in an interaction (*Penk, Sst, Gad2, Hoxb8*) that clustered with a glutamatergic annotation, its characteristic state is (1, 1, 0, 1), indicating that *Gad2* is indeed only a combinatorial marker for these glutamatergic neurons in its absence. Note that in 3 clusters (SCGLU4, SCGLU5, and SCGLU7), *Hoxb8* appeared in every interaction. Each of these three clusters corresponded to glutamatergic neurons of the spinal cord, but was defined by completely different neuropeptide genes: *Grp/Pthlh*, *Penk/Tac1/Sst*, and *Elfn1* (with the peptidase inhibitor *Serpine2*). In the Mouse Brain Atlas—the source of the expert annotations—none of these neuropeptides appeared as marker genes by differential expression (though the neuropeptide gene *Nmu* does appear as a marker for SCGLU5), and neither did *Hoxb8*. While the expert annotation thus found a similar number of states among the glutamatergic neurons, the characteristic states were able to associate to these particular sets of neuropeptides and highlight a role for *Hoxb8* that was not found by differential expression analysis.

The cluster defined by *Igfbpl1/Cd63/Id2/Tcf4/Tubb2b/Tac2/Sox11* (first non-singleton cluster from the bottom) clustered closely to two neuroblast annotations in the bottom of Figure 5.16, supporting the canonical role of *Sox11* and *Igfbpl1* in neuroblasts. Furthermore, that *Tac2* identifies a neuroblast subtype was already seen in [109]. More interesting were the cases in which the states did not closely map to a single expert annotation. For example, Figure 5.16 shows a polysemic state (in green) defined by two characteristic states involving *Synpr* and *C1ql2/3* that associated to a cluster of inhibitory neurons in the midbrain, as well as to granule neuroblasts in the dentate gyrus. Looking at its embedding in Figure 5.15 (fourth state from the left), it can indeed be seen that this state localised in two disjoint regions of expression space. *Synpr* encodes a synaptic vesicle membrane protein, *C1ql2/3* regulates the number of synapses, *Ybx1* is a transcription factor upstream of synapse development and signalling regulation [78]. This state thus corresponded to a cell state that is present in different cell types, which is likely involved in the regulation of synapse functioning and development.

Adolescent astrocytes Similarly to the adolescent neurons, I reclustered the characteristic states but included the binary representations of the expert annotations. This time the reclustering did not change any of the original clusters (AMI=1.0, ARI=1.0). The states that already had a high Youden index with respect to Bergmann glia indeed associated closely to that annotation in the dendrogram of Figure 5.16, most strongly so for the interaction involving the canonical Bergmann glia marker gene *Gdf10*. In contrast to the analysis of the Youden index, I found two more clusters of states that associated to a particular expert annotation, though at a larger Dice distance than was the case for the Bergmann glia case. The state that associated to the ACMB (*Dorsal midbrain Myoc-expressing astrocyte-like*) annotation included the genes *Slc38a1, Gria1* and *Fxyd6*, which are linked in Pathway commons as sharing a Reactome catalysis pathway. Furthermore, *Gria1* encodes a glutamate receptor subunit, *Fcyd6* encodes a glutamatergic synapse protein [29], and *Slc38a1* encodes a glutamine (a glutamate precursor) transporter. Beyond these glutamate-related genes, *Nnat* is involved in the regulation of ion channels during development [226], and *Antxr1* encodes another transmembrane protein. This state thus appeared to be related to active glutamatergic synapses. This is further corroborated by the fact that the strongest marker for the ACMB cluster in the Mouse Atlas was, after *Myoc*, the canonical astrocyte marker gene *Gfap*, whose promotor is targeted by Glutamate [221].

The ACNT2 (Non-telencephalon astrocytes, fibrous) associated state involved another gene that codes for the GFAP-interacting protein PLP1, that in turn interacts with the state-associated protein MBP [303]. Together, the MBP and PLP1 proteins account for around 80% of the myelin protein content in the central nervous system [303]. *Slc6a9* encodes a transporter protein for glycine, which stimulates myelin phagocytosis [48], as does the protein SCD1 [32]. Finally, *Itpkb* regulates the metabolism of inositol, a binding target of PLP1 [309] and one of the main constituents of myelin [247]. Together, these findings are consistent with the ACNT2 associated state corresponding to the control and metabolism of myelin. This is further supported by the fact that ACNT2 has the expert annotation of *fibrous* astrocytes, which are known to localise around myelated nerve fibres.

The authors of the expert annotation [313] used just the expression of *Gfap*—which canonically marks fibrous astrocytes—to annotate these cells. They only briefly commented on high *Slc6a9* levels, but did not link this to myelin phagocytosis, so this specific character of these cells was missed. That could be because not all ACNT2-annotated cells were in this state; this state only associated to ACNT2 at a relatively high Dice distance of around 0.8. This state annotation was thus not equivalent to the ACNT2 annotation and potentially included cells from different regions or types.

5.3.5 Semantics reproduce across data sets

Even though the states were robust to bootstrap resampling the cells before clustering the interactions, I wanted to verify that similar states would appear from a similar population even if the interaction estimates are based on completely disjoint sets of cells. That the interactions should be robust to this was already shown in Section 4.3.2, but I verified it here independently for the states.

I randomly sampled two disjoint subsets (DS) of 20,000 developmental neurons—referred to as DS1 and DS2—and ran the `Stator` pipeline on each cell population separately, estimating interactions among the 1,000 most highly variable genes of each data set. Because the genes in the analyses were different, individual interactions and states could not be compared directly, so instead I compared the biological semantics of the inferred states. The states that resulted from cutting the dendrogram at a Dice distance of 0.88 are shown in Figure 5.17. The *Six3/Gucy1a3+* D2 MSNs are found in both replicates, as are the *Isl1/Ebf1+* D1 MSNs. Furthermore, both replicates showed cycling states that separated into a G1S-phase state (marked by *Ccnd2* and *Cdca3*, respectively), and one *Cenpa+* that in DS1 seemed to correspond to a G2M state since both *Cenpa*, *Cdc20*, and *Ccnb1* specifically regulate the G2M phase in humans [297]. In DS2, the *Cenpa+* state was not clearly a G2M state as it involved the genes *Cdca3* and *Cdca8*, which have been implicated in the G1S as well as the G2M transitions [203, 27, 62], though especially *Cdca8* is canonically associated with mitosis, which suggests this state in DS2 might also correspond to the G2M transition.

More alarming were the states that appeared in only one of the replicates. For example, DS1 contained the *Hba/b* state seen before, but this state was absent in DS2. Similarly, DS2 contained a neuronal differentiation state (involving *Neurod2* and *Neurod6*, denoted Neurodx, and the neuron differentiation gene *Lmo1* [290]) and the neuronal activity state marked by the AP-1 genes *Fos/Jun+*. However, recall that a particular expression pattern of interacting genes was required to deviate more than 500% from the Bernoulli null to be considered a characteristic state. Upon closer inspection, this arbitrary threshold explained the difference in the found states. For example, while DS2 did not contain the Neurodx state, it did contain the 3-point interaction among *Neurod2, Neurod6* and the neuron differentiation gene *Id2* [103] with a characteristic $(1, 1, 1)$ state. However, this state was $\sim 450\%$ overrepresented, so is not shown as a state in the dendrogram. Similarly, while DS2 did not have the state with the four Globin genes, it had all three 3-point interactions between *Hbb-bs, Hbb-bt, Hba-a2*, and *Hba-a1* with the $(1, 1, 1)$ states $> 100\%$ overrepresented. Finally, the active neuron state marked by *Fos/Jun+* in DS2 is not found in DS1, but DS1 did contain a 3-point interaction among *Cyr61, Egr1*, and *Fos*—*Egr1* also being a marker for neuronal activity [71]—with a $> 200\%$ overrepresented $(1, 1, 1)$ state. The apparent irreproducibility of some of the states could thus be attributed by the stringent requirement of 500% overrepresentation. Therefore, I concluded that the semantics of the states were reproducible across the two replicates.

However, the concordance goes further than the states shown here. Setting the threshold for significantly deviating states at $> 500\%$ overrepresentation potentially hides more nuanced cell states. Because the states suggested that part of the population consisted of MSNs, I wanted to see if more of the lineage of this cell type was present in the population. MSNs and neurons from the olfactory bulb share a common precursor in the lateral ganglionic eminence (pLGE). The pLGE lineage splits into precursors of neurons in the olfactory bulb (OB potential), and cells with *striatal* potential, *i.e.* MSN precursors [240]. The MSN precursors then further differentiate into the D1 and D2 subtypes. The authors of [240] were able to identify two further subtypes of D1 MSNs in humans,

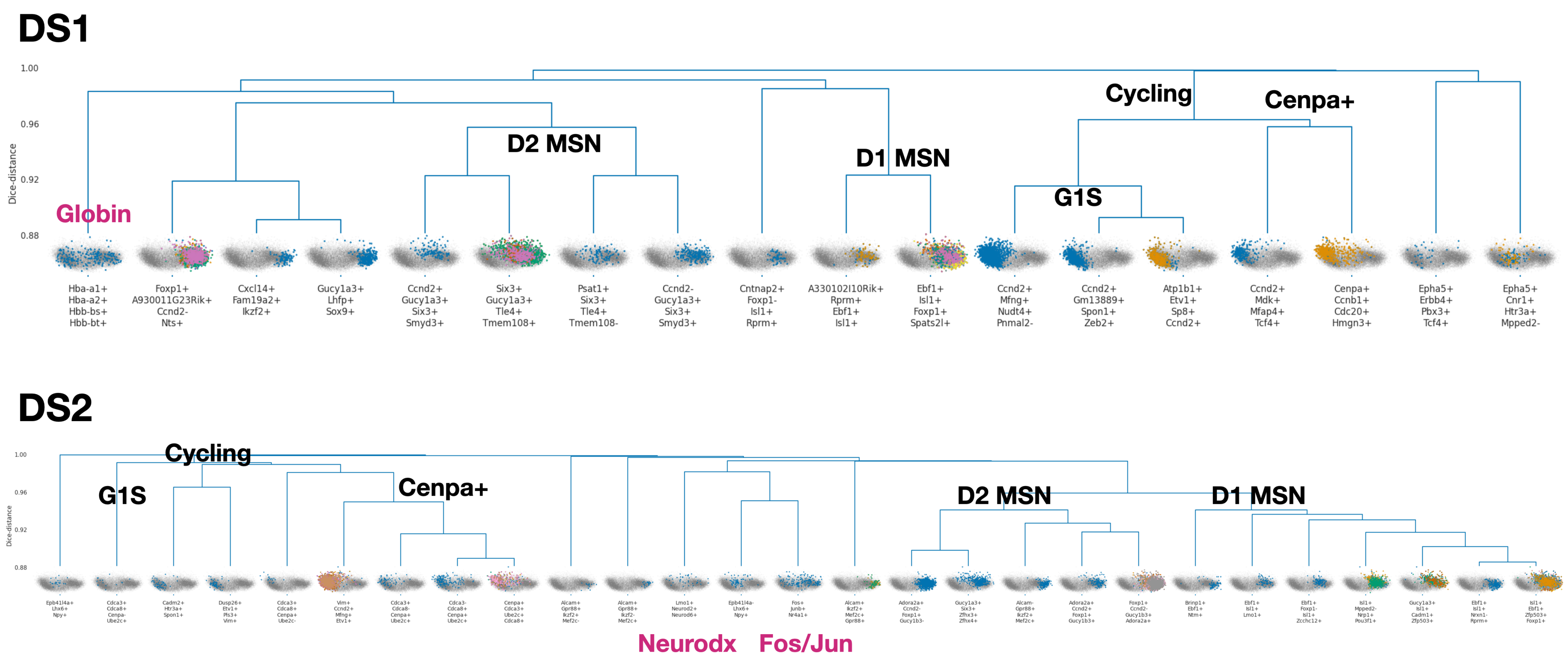


Figure 5.17: Two disjoint replicates from the develomental neurons resulted in mostly the same states. The states that were not reproduced (here in magenta) were still present in both data sets, but at a lower enrichment score.

marked by expression (Figure 4d and 4i of [240]) of the genes *Tac1, Ebf1, Isl1, Zfp503, Zfp521, Cntnap2*, referred to as *Pdyn* MSN precursors; and the genes *Tshz1, Foxp2, Pbx3, Erbb4*, referred to as *Tshz1* MSN precursors (all genes are mapped by human-mouse orthology). Furthermore, the authors of [240] further associated the genes *Chd7, Dlx5, Id2, Pax6, Dlx2, Arl4d* to the OB potential lineage, and the genes *Meis2, Tle4, Foxp1, Zfp503, Zfhx3/4, Id4, Bcl11b* to the striatal potential lineage. Figure 5.18 shows that both datasets contained states corresponding to each of the points along the pLGE lineage (ignoring D2 MSNs as those were already found in the dendrogram). It can be seen that the different lineages did not sharply localise in the PCA embedding, but rather showed a slight bias towards a region, though the dispersion pattern was similar across the two data sets. In both data sets (the left and the right column in Figure 5.18), the OB potential states all fell on the western blob of the PCA embedding, while the striatal potential spread out slightly east of centre (I used PCA embeddings for reproducibility, but the same structure was visible in tSNE or UMAP embeddings). The *Tshz1* MSNs tended to fall in the south-west and the *Pdyn* MSNs on the eastern blob. This meant that the purely data-driven approach that only found two significant clusters largely conflated the OB neuron progenitors with the *Tshz1* MSN precursors, and thus did not properly separate the MSNs from the other neurons. This also led me to conclude that this cluster of cells can be annotated as pLGE cells, not just CNS neurons as I did in Section 4.2.6.a. Finally, note that more than half of the interactions that defined the OB potential states involved the transcription factor *Tcf4*. This gene was not mentioned in [240] as a marker for the OB potential, but seemed to play a central role in the OB fate decision as it was part of the conditional expression pattern in this cell identity.

5.3.6 States can be inferred on data containing multiple cell types

5.3.6.a Polysemic states in neurons and astrocytes

Since the interactions could identify states not localised in expression space, one might ask whether the cells need to be separated by cell type, via data-driven clustering or otherwise, before the `Stator` analysis is run. This was indeed no longer a requirement, so I ran `Stator` on a merged data set that contained 10k of the developmental neurons and 10k of the developmental astrocytes. Cutting the resulting dendrogram again at a Dice distance of 0.88 led to a total of 40 states, the largest number of states so far. The full dendrogram split the states into two at a high Dice distance, but these two sides of the tree did not exactly correspond to separating the neurons from the astrocytes, with various polysemic states being present in both branches. Figure 5.19 shows five polysemic states highlighted, out of the 40 states in total. The left blob in the shown PCA embedding corresponded to the developmental neurons, and the right blob to the astrocytes. Immediately recognizable were the Globin+ state and the *Neurod2/6* neuron differentiation state, present throughout both cell types.

The leftmost highlighted state was composed of three characteristic states, all based on a negative interaction among *Fgf12/Pls3/Etv1*. ETV1 is known to interact with FGF proteins in particular during eye development [91, 299]. The triplet state $(1, 1, 1)$ was present in both cell types, but the triplet interaction was modified in a 4-point interaction with *Aldoc*. *Aldoc* is a canonical astrocyte marker, and here indeed unambiguously

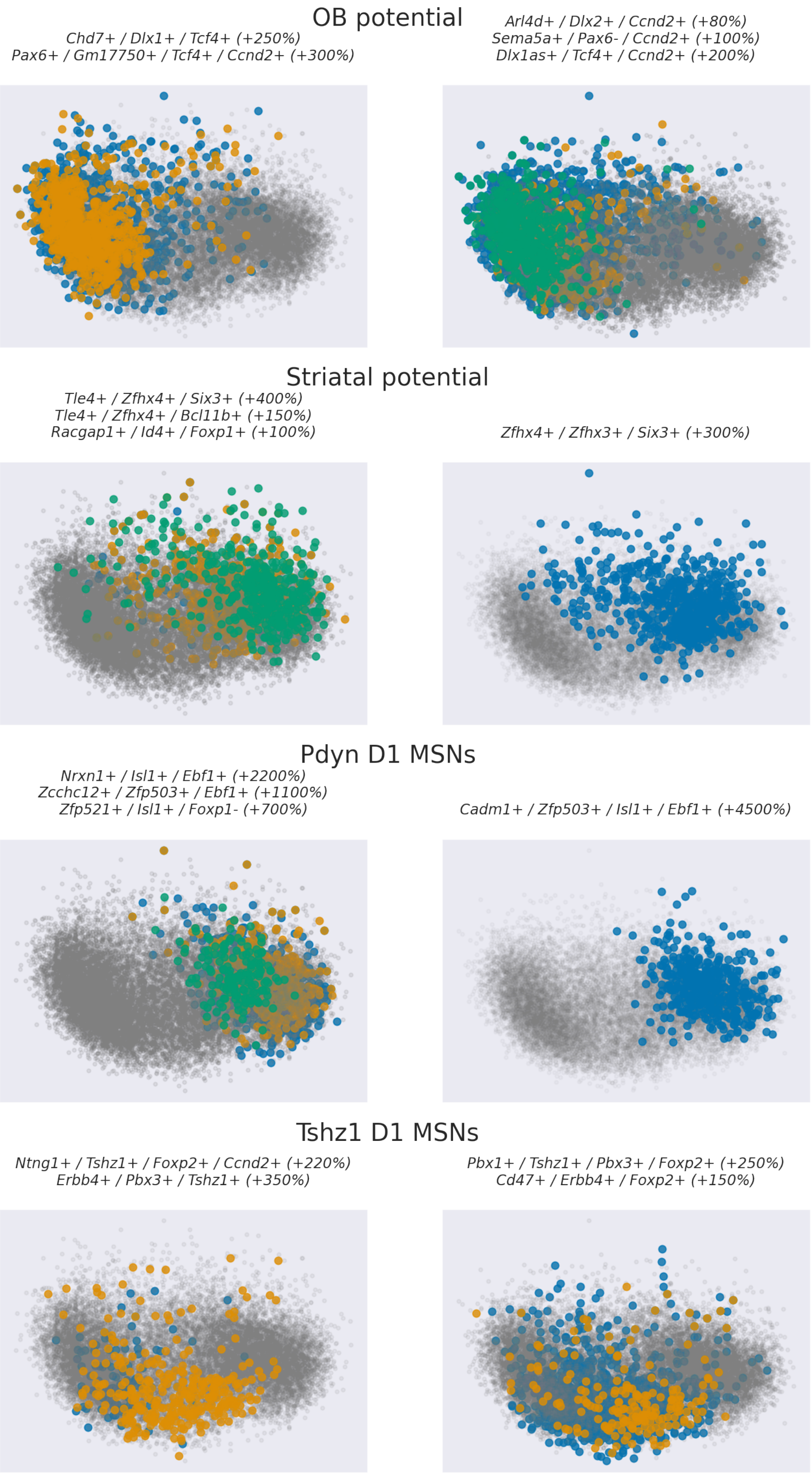


Figure 5.18: Four phases of the pLGE lineage all manifested as characteristic states of higher-order interactions, and showed similar localisation in DS1 (left PC embeddings) and DS2 (right embeddings). The overrepresentation relative to the Bernoulli null in the full data set is reported in parentheses, and each characteristic state that constitutes the full state is printed in a different colour. These characteristic states involved the marker genes from [240], and here indeed reproducibly localised in the same region, showing that even these fine-grained semantics—sometimes less than 500% overrepresented—were reproducible across data sets.

identified the astrocytes. Since the triplet interaction without *Aldoc* was negative and the 4-point interaction was positive, the presence of *Aldoc* lessened the absolute strength of the triplet interaction. The state thus appeared to correspond to an FGF-*Etv1* interaction that is strongest in neurons.

The fourth highlighted state consisted again of G2M cell-cycle transition genes, namely *Cenpa*, *Pbk*, and *Top2a*, but also the S- and G2-phase associated gene *2810417H13Rik/Pclaf* [155]. This state is discussed separately in Section 5.3.6.b. The last highlighted state centred around the genes *Fut9* and *Rxrg*. I could not find a biological interpretation for this state, though RXRG is known to physically bind to the protein LHFPL6 which is encoded by the *Lhfp* gene also expressed in the state (as reported by Pathway Commons, based on Supplementary Table 2 from [242]). This state was further marked by the absence of *Ebf1* expression, which could be considered surprising because *Ebf1* expression is known to be crucial to brain development [90].

Whereas most states were composed of just one or a few interactions (median 1, IQR 1-3), one state was composed of 175 interactions, *i.e.* 60% of all interactions with deviating states. Its highly dispersed embedding and the eight most commonly appearing genes are shown in Figure 5.20. All eight genes (except for *Mmd2*) were listed in Table 1 of [35] as core identity genes of neural stem cells (NSCs) and astrocytes in the hippocampal subgranular zone (SGZ). In fact, the top 4 occuring genes were expressed in at least 80% of SGZ NSCs and astrocytes in [35]. An NSC/astrocyte signature is typical of radial glia cells (RGCs), so this state might correspond to a general class of RGCs, especially because the markers for the SGZ—which develops in mice around postnatal day 7—are unlikely to be useful in embryonic data. However, the large number of interactions and genes that constitute this state make a full interpretation difficult, and cutting the dendrogram at a lower Dice distance might reveal more nuanced structure within this state.

5.3.6.b Combinatorial markers for the cell-cycle

One interaction in particular stood out because it had 5 states that passed the 500% overrepresentation threshold: The 4-point interaction among *Top2a*, *Pbk*, *Cenpa*, and *2810417H13Rik/Pclaf*. It is also noteworthy that the STRING database links all these genes together with high confidence (> 0.7), based in part of direct experimental verification, which means that this set of four genes is enriched in interactions with a p-value of $5.4 \cdot 10^{-7}$ relative to the STRING database's null hypothesis based on a random set of proteins of the same size and degree distribution drawn from the full mouse genome (for four genes, this amounts to zero expected interactions, so the enrichment factor is undefined). The $\widehat{1}$-state was the most deviating state at over 15,000% deviation, but it occurred about as often as the four states where 3/4 genes were expressed, all of which passed the threshold of a significant $>$ 500% deviation. A summary of the various states is shown in the UpSet plot in Figure 5.21, and an embedding of the 5 characteristic states is shown in Figure 5.22.

The authors of [230] have annotated a population of mouse brain cells from a variety of public data sets by cell cycle phase. An embedding of these cells and their annotation in the UCSC Cell Browser [252] is shown in Figure 5.23. It can be seen that *Pclaf* was

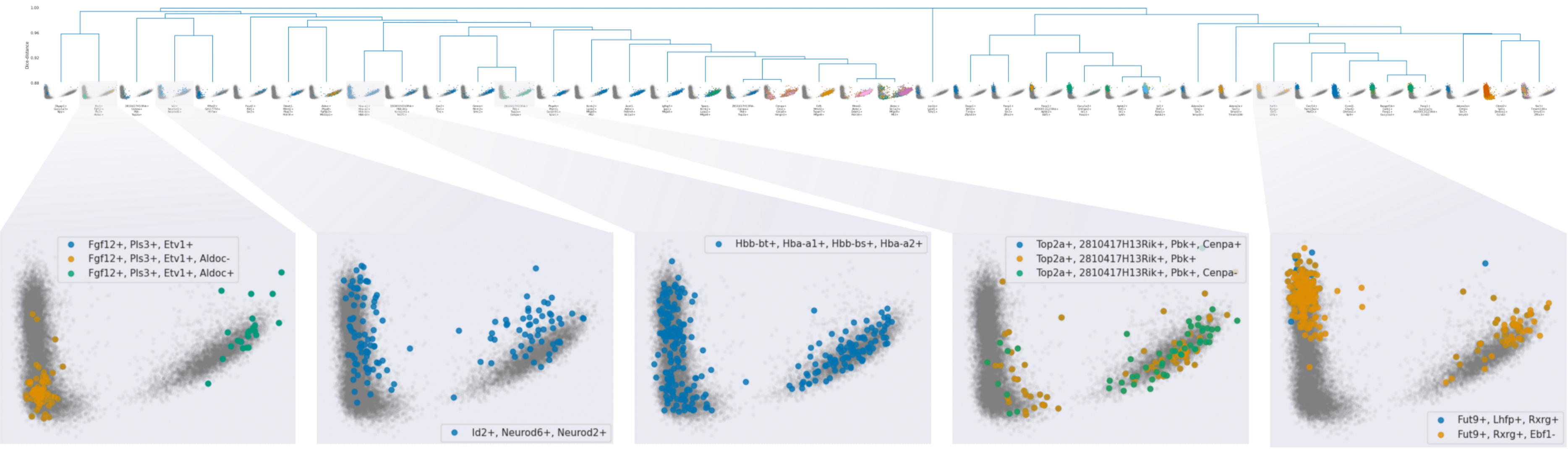


Figure 5.19: The merged data set led to a total of 40 states after cutting the dendrogram at a Dice distance of 0.88. Here, five polysemic states are highlighted that were present in both the astrocytes (left blob in PCA space) and neurons (right blob). From left to right: A cell type dependent FGF-ETV1 interaction state; a neuronal differentiation state; a *Hba/b* expression state; a G2 phase cell state; a poorly understood state.

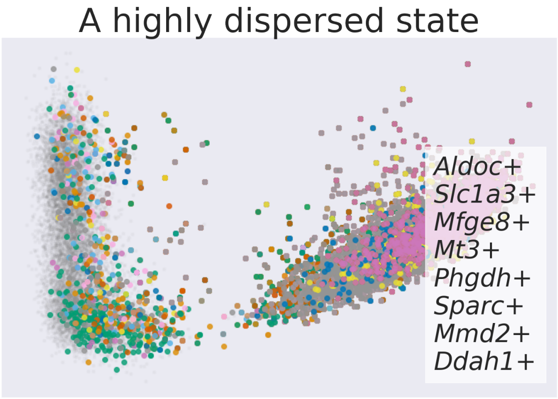


Figure 5.20: A state composed of 175 interactions, with the eight most commonly occurring genes listed. These eight genes suggest these might be radial glia cells, but the large number of interactions that constitute this state makes interpretation difficult.

primarily expressed in the S-phase, and decays afterwards, *Pbk* and *Top2a* were expressed in the early- to mid-G2M cluster, whereas *Cenpa* sharply marked the late stages of the G2M annotation. Note that these genes mostly marked the phases by the presence of their transcripts, not the absence, but it can still be seen that the absence of only *Cenpa* corresponded to the S-phase, and the absence of only *Top2a* to either late G2M transition or the mitotic exit. This interaction therefore seems to be the result of cells being at various stages of the cell-cycle throughout the population, marking the different phases by the different combinations of genes expressed.

5.3.7 Assigning bootstrap significance to branches

Up until this point, I defined cell states as collections of characteristic states that looked 'similar enough', *i.e.* clustered by Dice distance. This had multiple drawbacks. First, the choice of threshold can strongly affect which and how many states are inferred. Setting the threshold lower than 0.88 can lead to many more states, but is less conservative as the different states might then correspond to identical biology. At the same time, a high threshold conflates many characteristic states, potentially destroying a lot of information and structure in the data. This was already seen in the previous section where a state consisting of 60% of all interactions emerged. Finally, the states were robust in terms of the AMI and ARI of bootstrapped reclusterings, but these quantities only describe how robust the clustering is *on average*, and say nothing about the robustness of individual states. To overcome these challenges, I switched to quantifying the robustness of the clustering by assigning bootstrap confidence values (the BP and AU score, see Section 5.2.2) to the individual branchings.

Figure 5.24 shows one of the stable branches that resulted from running `Stator` on the merged data set of developmental neurons and astrocytes, where stable was defined as having an AU value larger than 0.95 (and significant values are marked in green). The full branch is indicated with the letter A, and was robust with AU=0.97 (with a BP of 0.94). Branch A contained many *Six/Sp9/Foxp1/Gucy1a3+* states, all markers for MSNs or the striatal potential. This set of states thus seemed to correspond, again, to the striatal branch of the pLGE lineage. In particular, the *Gucy1[x]3* genes and *Sp9*

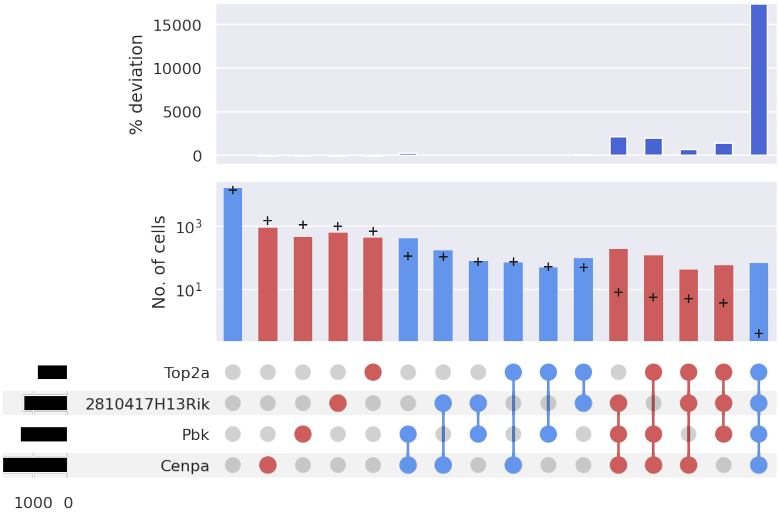


Figure 5.21: This UpSet plot shows the frequency of the various expression patterns associated to the 4-point interaction *Top2a/Pbk/Cenpa/2810417H13Rik(Pclaf)*. The frequency of each of the $2^4 = 16$ states that appear in a 4-point interaction is shown, as well as the % deviation from the Bernoulli null. The frequency is shown in blue if the state is part of the numerator, in red if it is part of the denominator. The black bars indicate the marginal expression level of each of the four genes. The $(1, 1, 1, 1)$ state deviates the most, but all three states where one of the genes is not expressed also pass the $>$ 500% deviation threshold and are deemed characteristic states of this 4-point interactions.

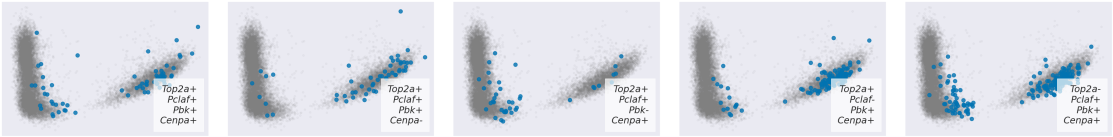
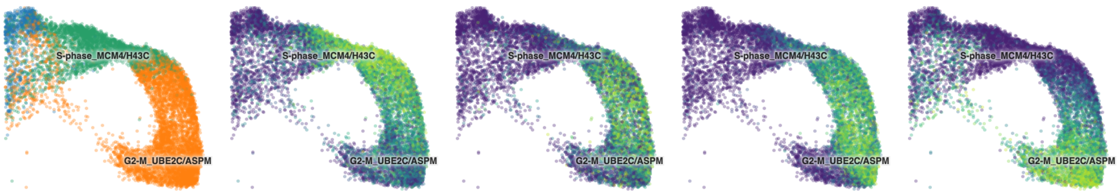


Figure 5.22: The five polysemic states, present in both neurons and astrocytes, associated to the interaction *Top2a/Pbk/Cenpa/2810417H13Rik(Pclaf)*.



(a) UCSC clustering (b) *Pclaf* expression (c) *Pbk* expression (d) *Top2a* expression (e) *Cenpa* expression

Figure 5.23: The cell-cycle annotation in the UCSC Cell browser showed that expression of the four interacting genes marked distinct phases of the cell cycle (here shown for embryonic mouse, but results were similar for adult mice).

mark D2 MSN progenitors. Note, however, that the branches of the state dendrogram do not need to correspond to branches of the developmental lineage.

First a singlet (the leftmost state) split off from the rest of the states, but it did not do so robustly (AU$= 0.81 < 0.95$). Therefore, the set of cells that it identifies cannot be seen as different from cells in the other states. More interesting was the robust split into the branches labelled B and C. Branch C was a *Cited2/Sp9/Dlx6os1+* cluster. *Dlx6os1*, a lncRNA antisense to *Dlx6*, has been shown to regulate GABAergic fate decision of progenitor cells in the ganglionic eminences [36] which is in line with the pLGE MSN lineage. Branch C itself had two robust (unlabelled) sub-branches—one with *Ccnd2–* and one that includes a *Ccnd2+* state—that therefore appeared to correspond to two distinct points along the cell cycle.

Branch B was a more diverse branch than branch C and split into three robust branches labelled D, E, and F. Branch D was a *Six3/Smyd3+* branch, and although it is not obvious why *Smyd3* appeared in these interactions, the fact that it showed up in an RNA-based investigation is not surprising as it encodes a chromatin methylating protein that is part of the RNA polymerase II complex [102], and is thus directly involved in the process of transcription. Branch D itself had two (unlabelled) robust branches with *Adora2a+/–* respectively. *Adora2a* forms a heterodimer with the dopamine 2 receptor, encoded by the canonical D2 MSN marker gene *Drd2* [45].

Branch F mainly centred around the gene *Rasgef1b*, that plays a role in GTPase mediated signal transduction. It appeared together with *A930011G23Rik*, but these two genes are less than 50kb apart on chromosome 5, so might be conditionally expressed simply due to their proximity. This branch also featured *Gucy1a3*, a gene that encodes part of the soluble guanylyl cyclase enzyme which converts GTP to cGMP and hence also plays a role in GTP/GDP signal transduction [53]. Finally, the role of *Calb1* here was not directly clear, though it is known to bind directly to the GTP regulator *Ranbp9* [154].

Branch E contained a robust branch around the *Zfhx3* gene, which is enriched in the striatal potential lineage [240]. The transcription factor *Sp9* directly binds to the promoter and a putative enhancer of *Six3* [308], in particular during D2 MSN development [249], but although both genes appeared in this branch, they did not appear in the same interaction.

Similar analysis of other bootstrap robust branches revealed that the full dendrogram first split into three distinct branches: Neurons, Astrocytes, and a range of polysemic states, after which the neurons robustly separated into the D1 and D2 MSNs. I did not find evidence for the various D1 subtypes found in Section 5.3.5, but this could be explained by the fact this data set included only half (10k) of the neurons in the data sets from Section 5.3.5. However, there were many robust sub-branches that were hard to interpret. How to extract the hierarchy of meaningful states is an open problem, and briefly addressed in Section 6.2.

5.3.8 No evidence for interactions beyond fifth-order

There are two reasons to expect to find significant interactions only up to a certain order. Biologically, a true 6- or 7-point interaction that cannot be decomposed would correspond

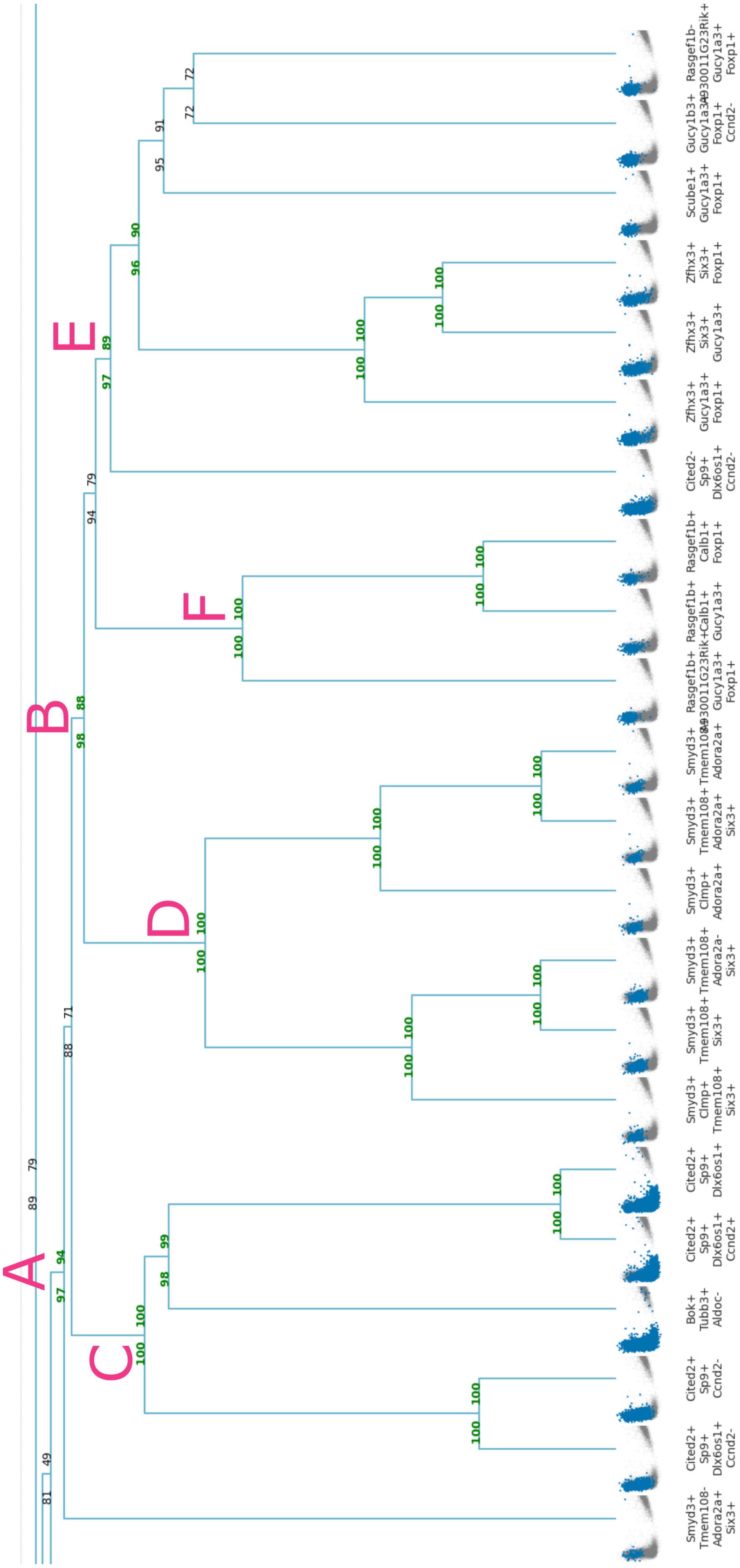


Figure 5.24: One of the stable branches from the bootstrap resampled dendrogram with states in the merged data set of developmental neurons and astrocytes. Reported values at each branch are AU (left) and BP (right). Note that the vertical axis still corresponds to the Dice-distance, but stretches from 0 to 1, *i.e.* the dendrogram is not cut.

to highly complex regulation that requires the coordination of many molecules. That cells are able to coordinate such an interaction using the inherently stochastic process of transcription in a noisy environment can be considered unlikely in the absence of direct evidence. However, even if there are true dependencies at such high orders, the number of cells that go into each part of the estimation will become smaller and smaller, hiding all dependencies beyond a certain order in sampling noise. Nonetheless, the existence of 6- and 7- point interactions cannot be excluded. However, calculating all 6- and 7-point interactions, or even only the Markov-connected ones, would take prohibitively long because there are simply too many to calculate. Therefore, I used two different approaches outlined below.

Random tuples In each of the four data sets, I calculated 6- and 7-point interactions between randomly chosen 6- and 7-tuples. For each of these, I calculated the interaction using the Markov blanket of the first gene. Of these in total 4k interaction estimations, each one involved states that were not present in any of the cells, which means that none of these interactions were estimable.

Targeted search Alternatively, one might hypothesise that the presence of an estimable and significant 5-point interactions might increase the probability of further higher-order interactions within the shared Markov blanket of the interacting genes. In each data set, for each significant 5-point interaction already found, I calculated 6- and 7-point interactions between 6- and 7-tuples of fully Markov connected genes that include the 5 interacting genes. The results are shown in Table 5.4. It can be seen that the only interaction that was perfectly estimable and significant at $\alpha = 0.05$ is a 6-point interaction in the adolescent neurons. This was the 6-point interaction among the already interacting pentuplet *Slu7, Trim35, Ube2n, Lmbrd1, Vsnl1*, and the gene *Plp1*, with a point estimate of -2.54, a 95% confidence interval $(-4.80, -0.35)$, and an F-value of 0.013. Looking at Figure 5.16, the 5-point interaction was already closely related to other interactions that involved *Plp1*, so this 6-point interaction did not add much new information. Across the 4 data sets, no 7-point interaction was perfectly estimable, and even among the interactions that could not be perfectly estimated, the smallest F-value reached was 0.20. I thus concluded that there is no evidence for beyond fifth-order interactions in these four data sets.

5.4 Discussion

In this chapter, I associated characteristic states to the higher-order interactions from various data sets, and investigated their biological meaning. To do this, I created a publicly accessible `Nextflow` pipeline called `Stator` that estimates all 1- and 2-point interactions, and all 3-, 4-, and 5-point interactions among Markov-connected tuples of genes. Sufficiently deviating states among the higher-order interactions were deemed characteristic states, and upon being clustered proved to reflect biological cell states.

This process is fundamentally different from assigning cell identity by annotating clusters in expression space, which has the disadvantage that it can only justify clusters by a difference in mean expression. Furthermore, usually the same cells that led to the

	Developmental		Adolescent	
6-points	Neurons	Astrocytes	Neurons	Astrocytes
Total	10	16	26	2
Estimable	2	9	9	2
Perfectly est.	0	2	2	1
Sig. $\alpha = 0.1$	0	1	2	1
Sig. $\alpha = 0.05$	0	0	1	0
7-points				
Total	8	13	21	0
Estimable	0	1	3	0
Perfectly est.	0	0	0	0
Sig. $\alpha = 0.1$	0	0	0	0

Table 5.4: Shown is the fraction of 6- and 7-point interactions (among genes with at least one 5-point interaction) that were estimable and significant. An interaction is estimable if its point estimate is estimable, and it is perfectly estimable if all bootstrap resamples were estimable.

clustering are used in the differential gene expression analysis. This process is called selective inference and is known to lead to false conclusions. Since the interactions and their characteristic states are defined without making reference to distances in expression space, they do not suffer from selective inference bias.

The inferred states indeed revealed structure beyond what can be accounted for by clustering cells in expression space, and revealed different cell identities from the expert annotation, where that was available. Furthermore, the states were robust to bootstrap resampling the cells before clustering, and even to estimating the interaction on a completely disjoint set of cells from the same population. Some states did not localise in expression space, and were thus *by definition* invisible to clustering cells in expression space. I called such states *polysemic* states, as they comprise a consistent higher-order gene expression pattern across multiple cell identities.

In a population of developmental neurons, annotated as a single cluster by *10X Genomics*, further clustering by expression could split this cluster in two, but the characteristic states identified many more states. In particular, the states revealed that the cells were not just CNS neurons, as a differential gene expression analysis suggested, but in fact corresponded to a population of progenitor cells originating from the lateral ganglionic eminence. The states corresponded to various branches from this lineage, like those with an olfactory bulb fate, and those with a striatal fate, which separated into progenitors of D1 and D2 medium spiny neurons (MSNs). It was surprising that these two cell types emerged so clearly, since the two genes that canonically define these states *Drd1* and *Drd2*, were expressed very lowly or not at all. However, this is known to be the case in developing MSNs, where the regulatory markers are present before *Drd1/2* are expressed [319]. I further found two subtypes of D1 medium spiny neurons: *Pdyn* MSNs, and *Tshz1* MSNs. The *Pdyn* subtype is also marked by the *Tac1* gene, which is how this subtype was found by [240] in humans. A *Tac3*-marked D1 MSN subtype was found in Rhesus monkeys in [104], and [320] found that *Tac2*-expressing MSNs played a role in a rodent model of cocaine addiction. Recently, similar subtypes have been described in postnatal mice in a data-driven [313] and targeted [307] investigation of *Tshz1+* cells. The authors of [253] describe a data-driven *Tac1+/Penk+* D2 MSN subtype in mice, but as this is a D2 subtype it seems to correspond to a different type from that of [240]. I therefore believe that this is the first time the *Pdyn* and *Tshz1* D1 MSN subtypes have been found in embryonic mice in a purely data-driven way. Furthermore, the *Pdyn* subtype was identified by the genes from the genetic programme underlying the differentiation as listed in [240], not by *Pydn* directly.

Beyond these two cell subtypes, I found states corresponding to distinct phases of the mitotic cell-cycle, p53/DREAM-regulation, myelin metabolism, *Arc*-capsid transport, *Dlk1*-imprinting, and neuropeptide signalling. The cell-cycle states were often polysemic and present in multiple cell identities. The *Dlk1*-imprinting state seemed to correspond to a neurogenesis state among niche astrocytes or radial glial cells. However, these neurogenic niches only become active and generate neurons in adulthood (at which point they are the main source of neurogenesis), whereas the data only comprised embryonic cells. It is known, however, that late-embryonic radial glial cells already generate the neural stem cells (NSCs) that are responsible for adult neurogenesis [164, 157]. Since the cells were indeed from late-embryonic (E18.5) mice, the found state likely corresponded to NSCs derived from radial glia cells. Note that no evidence of neurogenesis was found in the differential expression analysis for the whole population, which only showed an astrocytic marker signal. That neurogenic radial glial cells share many features—including markers genes—with astrocytes is a well-known fact that has led to confusion about cell identities before [180]. If this state indeed corresponds to a precursor of adult neurogenesis, that would imply that the *Dlk1* imprinting already happens before adult neurogenesis starts. While the authors of [80] found no maternally expressed *Dlk1* at E14.5, that the maternal allele becomes active around birth could be possible and investigated in future research.

Multiple data sets also showed further polysemic states corresponding to the expression of Globin genes and neuronal activity. The Globin genes expressing state could be the result of doublet transcriptomes that include a blood cell, but if that were the case, one would expect the cells in that state to separate from the neurons and astrocytes and colocalise with similar doublets in expression space. Since this is not seen and the cells in the Globin state are present throughout expression space, it most likely corresponds to a cell state that both the neurons and astrocytes could enter. Such neurons and astrocytes have indeed been found in both mice and humans [28].

The discovery of these polysemic states suggested that the cells did not need to be separated by type before inferring the interaction and states. I ran the same analysis on a data set containing both neurons and astrocytes, and found a very similar set of states, with the polysemic states now present throughout both cell types. While the statistical power within each cell types decreased due to a smaller number of cells of each type, the added heterogeneity in fact helps with the estimation of the interactions, as a more diverse set of gene expression patterns become available. This suggests that estimating the interactions and states on an even more diverse set of cells—perhaps even an organism-wide data set—might also prove tractable and interesting.

Throughout this section, the validation of the states has gone in one direction: I have annotated the states by comparing the interacting genes with descriptions from the literature. An obvious next step, and even stronger validation, would be to turn this around and come up with a testable hypothesis about an inferred cell state, and validate this in my data. For example, the DREAM-complex binds not only to E2F binding domains, but also to CHR sites. Verifying that the genes involved in the DREAM-regulated states contain such sites would provide further validation of such regulatory states. Alternatively, a particular interaction could lead to predictions about the cells in its characteristic state, which could then be validated in independent data sets. For example, I found the gene *Hoxb8* to be involved in many of the glutamatergic neuropeptide signalling states, which could be verified in other such data sets. I also found that the olfactory bulb potential lineage in the developmental neurons contained multiple interactions involving *Tcf4*. Verifying that *Tcf4* plays a role in the OB fate decision in other data sets as well would amount to a strong validation of both the states and the interactions, and provide new biological insight into that lineage. While I could find no direct evidence for this yet, *Tcf4* does plays a role in the OB lineage of oligodendrocytes progenitor cells [317], and in neuronal maturation and differentiation in the hippocampus and anterior commissure of the embryonic mouse brain [165], both of which are functionally and anatomically related to the olfactory bulb.

Finally, all results in this chapter were based on interactions at up to fifth-order, as I found no evidence for 6- or 7-point interactions. However, this absence of evidence is not evidence for absence, because almost all 6- and 7-point interactions were inestimable (see Table 5.4), rather than zero. This might be because the sample sizes were simply too low, in which case such beyond fifth-order interactions could perhaps be found when estimating interactions on transcriptomic data sets containing more cells, which nowadays are commonly available. Alternatively, since the heterogeneity required to make an *n*-point interaction estimable increases exponentially with *n*, the transcriptome might simply be too homogeneous to estimate the interactions, regardless of sample size. This second option could pose a tougher challenge to overcome, but conditioning on different states of the Markov blanket might improve the relevant statistical power.

Chapter 6

Discussion

Controlled or not controlled?
The same dice shows two faces.
Not controlled or controlled,
Both are a grievous error.

Wumen [214]

䷅ ䷗

6.1 Limitations & further remarks

In this thesis, I constructed networks of interactions among genes based solely on measurements of RNA concentrations—a *mono-omic* approach. Such RNA measurements are inherently noisy, and do not reflect the full state of the cell. To properly understand the dynamics of genes and their products, a more integrated *multi-omic* approach might be better suited and provide a more holistic view on the inner workings of the cell. New single-cell technologies are being developed at a dizzying pace, and over the course of the last four years—the time it took to finish this PhD—new *dual-omic* technologies have emerged that combine simultaneous measurements of RNA molecules and chromatin accessibility. A data set of simultaneous measurements of a cell's transcriptome (scRNA-seq), proteome (single cell proteomics like CyTOF [272]), and chromatin accessibility (scATAC-seq [145, 227]) would provide a full view of the central dogma, and its dynamical extension in Figure 1.1, in action. While such multi-omic data sets would probably trigger the development of completely new analysis techniques, they could also be integrated into a single data set and analysed with the `Stator` pipeline introduced in this thesis. Genome-wide single-cell proteomics in particular is challenging due to the importance of a protein's three-dimensional structure, but since proteins are particularly stable and abundant in cells they might provide a less noisy measurement of cell state. As new technologies emerged, yesterday's innovations became easier, cheaper, and more readily available. When I started this research, the *10X Genomics* million cell data set was the largest data set of single-cell RNA-seq available, but many more have since become publicly available. These are all inherently interesting to study using the 'big data' approaches taken in this thesis, but one of the surprising results of this thesis was that a few thousand cells sufficed to give a deep and novel view into a cell population's structure.

Another surprising result was that binarising expression data did not seem to hide much of the biology. In the binarised data, I found biological structure that had not been found on unbinarised data before. This has been previously observed in other studies, and is usually attributed to the fact that scRNA-seq measurements are so noisy that it is accurately represented as binary data. As very deep sequencing becomes cheaper and more commonplace, it is not obvious that this will still be the case. At the very least, binarising around 0 and 1 transcript counts will not be sophisticated enough any more, so new approaches will have to be explored.

A limiting factor in this research was computational efficiency. Running `Stator` on 1,000 genes across 20,000 cells took around 5 days on the Eddie cluster, but ran in parallel on multiple cores so took around 2,500 CPU hours (as reported by the `Nextflow` pipeline manager). As further speed-ups are implemented (like the one discussed in Section 6.2.1.a), this might become more efficient, but scaling it up to a genome-wide analysis, *i.e.* an order of magnitude more genes, might prove prohibitively complex. The actual complexity of the full estimation also depends on the biology of the cells under study, as more complex biological systems might have larger Markov blankets. However, even in biological systems like the adolescent astrocytes studied in this thesis, that superficially seem to only contain 'simple' biology (as measured by sparse correlations and many housekeeping genes), I found interesting structure hidden in the characteristic states.

Finally, current RNA-seq methods are limited in the sense that they are destructive: the cells have to be destroyed to be sequenced. This makes it impossible to see the expression dynamics directly, which can distort the results and conclusions. For example, I found immediate early genes to be strongly represented among the higher-order interactions, but it is not clear if they would have been active in non-dissociated cells. Time-series data on genome-wide expression in single cells would truly reveal the dynamics and offer incredible insight into the cell's inner workings, but require new analysis techniques.

6.2 Future work

This research project raised at least as many questions as it answered, and provided multiple starting points for future projects. On a practical level, there are some obvious improvements and extensions to the methods used and introduced in this thesis. For example, many of the validation methods in Chapter 4 were currently only explored for pairwise networks, but might be extended to higher-order methods. In particular semantic similarity can be readily generalised to quantify the similarity between triplets of genes. Other notions of semantics might also be explored, like the phenotypic semantics in [112]. In addition, there are many graph-theoretical quantities that have a natural generalisation to hypergraphs, which could offer more insight into the structure of the interactions as a whole, rather than analysing the orders separately. For example, it would be interesting to compare a gene's function with quantities like node centrality and betweenness. The modularity of the full graph could also be explored more deeply, as well as other global quantities like the Laplacian spectrum and the algebraic connectivity. All these methods have provided insight into the structure of pairwise biological networks, and it would be valuable to calculate these here as well.

While the current method is completely data-driven, there are numerous ways to integrate biological knowledge into the estimation procedure that might make the results more accurate, meaningful, or relevant. For example, while I used causal discovery techniques to construct the Markov blankets, one might consider adding causal edges between genes whose proteins are known to interact, making the quasi-causal graph biologically more accurate. In addition, I only conditioned on all genes in the Markov blanket being unexpressed. This made the connection to Ising models and information theory most straightforward, but did not necessarily correspond to the most biologically plausible situation, and probably failed to extract all regulatory information. Conditioning on certain genes in the Markov blanket being expressed rather than unexpressed might reveal completely new interactions that are only present in a particular cellular context. Note that such interactions naturally arose in Chapter 3, where they were called outeractions. This approach is already being explored, and all relevant functionality is part of the `Stator` utilities package.

One can wonder about other questions the higher-order states could address. One particularly appealing idea is to construct a 'state atlas' across a whole organism, similarly to how that is currently done for cell types. Running the pipeline on all tissues from an organism—either separately or on one combined data set—could provide a very wide and deep view on the cellular dynamics across tissues. However, since states are dynamic and relate to a cell's environment, one might have to create such a state atlas from organisms across a wide range of experimental and physical conditions. It is unclear how large the state space of cells is—it might even be unbounded—so a state atlas would be a very interesting but ambitious project.

Finally, one could also move beyond cell states, to other biological identities. For example, if `Stator` runs on epidemiological data, then the resulting states describe coherent 'states' within the cohort, *i.e.* groups of people that share a group-identity through a non-random combination of traits or comorbidities. Alternatively, states inferred on mutational data from different cancer cell lines would reveal 'mutational' states that reveal which mutations preferentially come together, and which cancers share such states. More different still, one could expect a wide variety of species, types, and states within a population of single-celled organisms. In fact, even species are often not well-defined in such contexts, as horizontal gene transfer blurs the line between individual genomic identities. An RNA-based approach might offer new insight here, assigning identities based on transient transcriptional states based on a reference 'genome' that included all genes present in a population without making reference to any particular species' genome.

6.2.1 Tricks to improve statistics

The method as it is described in this thesis is still under development and there are numerous ways in which it can be improved or extended. A number of these extensions are already implemented and briefly described here.

6.2.1.a Constructing the asymptotic MFI-distribution

The MFIs are in essence conditional log-odds ratios. Odds ratios are generally assigned confidence intervals using various bootstrapping methods [217], but log-odds ratios can be analysed using the so-called *delta method* for error propagation to construct the asymptotic distribution of a statistic [6]. The delta method relates the variance of a statistic to the variance of some function of that statistic. In particular, if the original statistic is normally distributed, then the transformed statistic is also normally distributed, with a variance determined by the Jacobian of the transformation. In practice, this means that an n-point MFI has a standard error given by

$$\mathrm{SE}(I_{1\dots n}) = \sqrt{\sum_{i=1}^{2^n} \frac{1}{n_i}} \tag{6.1}$$

where n_i is the number of cells in the ith state. I have briefly motivated this method in Appendix 6.A, where I also showed that Equation (6.1) accurately describes the variance of the estimator and thus that the variance depends only on the probabilities of the different states. Using this asymptotic approximation for the F-value eliminates the need for bootstrap resampling, and could thus significantly speed up the estimation procedure.

6.2.1.b Dealing with divergent interactions

So far, I only considered interactions that were perfectly estimable, which means that the point estimate and all bootstrap resamples were well-defined. This need not be the

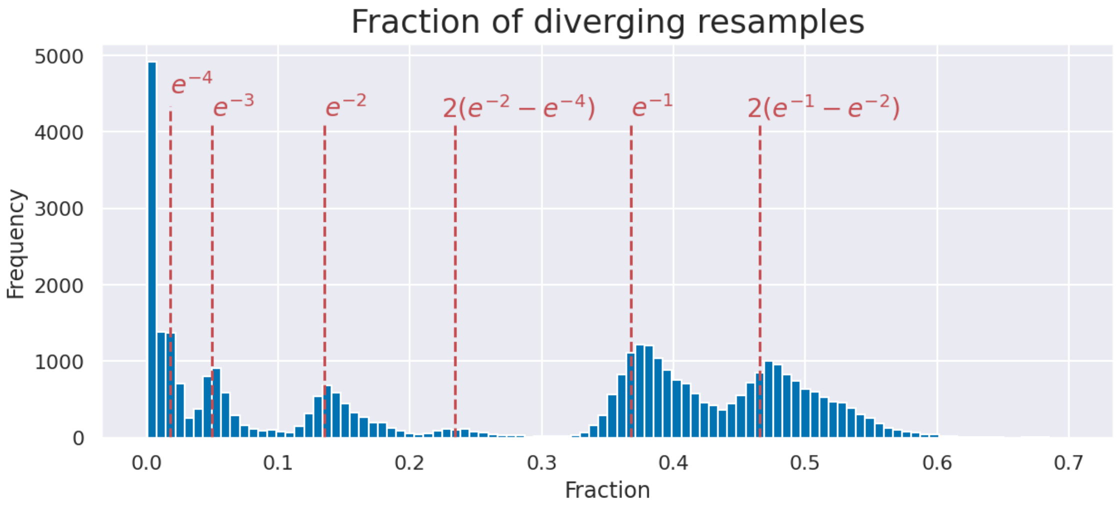


Figure 6.1: The histogram of the fraction of diverging resamples (here shown for 3-point MFI estimations in the developmental astrocytes) has multiple peaks. The peaks at a fraction of e^{-n} corresponds to the situation in which the smallest bin was of size n and all were omitted from the bootstrap resampled data set. When there are two states with only one cell—one in the numerator and one in the denominator—then *either* has to be omitted to make the estimate diverge, but not both (then the interaction would be undefined). The probability of this happening is $2e^{-1}(1-e^{-1})$, which explains peak at $2(e^{-1}-e^{-2})$. The long tail of this peak might be the result of it overlapping with the peak (not shown) that corresponds to one state with just one cell, and one states with two cells, that has probability $3e^{-1}-5e^{-2}+2e^{-3}\approx 0.53$. The small peak at $2(e^{-2}-e^{-4})$ corresponds to the situation where the two smallest bins are both of size 2.

case. Upon resampling with replacement, as is done in bootstrap resampling, it could happen that the cells needed to make the interaction estimable are omitted. In fact, the probability that any particular cell does not end up in a resampled data set is $(1-\frac{1}{N})^N$, where N is the total size of the resampled dataset, which for large N tends to $1/e$: $\lim_{N\to\infty}(1-\frac{1}{N})^N=\frac{1}{e}$. If any of the probabilities in an interaction were estimated from a single cell, then any bootstrap resample will result in an undefined interaction with a probability that tends to e^{-1}. Looking at the histogram of undefined resamples in Figure 6.1, we indeed see a peak around this fraction, as well as peaks that correspond to the various other situations that can lead to divergent estimates. If such divergent estimations can be dealt with, more interactions might become estimable.

One option is to make the estimator a map to the *extended reals* $\overline{\mathbb{R}}=\mathbb{R}\cup\{-\infty,+\infty\}$ by recognising that $\lim_{\epsilon\to 0}\log(\epsilon)=-\infty$ and $\lim_{\epsilon\to 0}\log(\epsilon^{-1})=+\infty$. As long as the original estimate did not diverge, estimates with divergent resampled estimates might still be informative. In fact, a 95% confidence interval could still be finite as long as fewer than 2.5% of bootstrap resamples diverge. Moreover, even infinite confidence intervals that stretch from a finite value to $\pm\infty$ can be informative and reveal an interaction to be significantly positive or negative. By assigning the value $+\infty$ to any bootstrap resample where only probabilities in the numerator are inestimable, and the value $-\infty$ when only probabilities in the denominator are inestimable, significance levels and confidence intervals can be calculated as before.

This technique only works as long as the initial estimate on the non-resampled data was well-defined and finite. If the original estimate already diverges, then all resamples will as well, and the confidence interval or significance cannot be properly estimated. One solution is to note that adding one cell to an empty bin might make the interaction estimable. If the added cell ends up in the numerator, then the resulting estimate forms an upper bound on the true estimate, and if the added cell is used in the denominator the estimate forms a lower bound. This allows even interactions that always diverge to be bounded. However, whether the confidence intervals associated to this bound are consistent with those of the true interactions is unclear, so further research on this is necessary before this method can be used to put bounds on interactions.

Both these methods of dealing with divergences are implemented in the *Stator* pipeline but not used in this thesis.

6.2.1.c Duplicating the data and compensating for decreased variance

Since it is only the *relative* proportions of the various cell states that determine the value of an interaction, the data can be freely duplicated without changing the value of the interaction. Since the variance of an estimator depends on the number of samples, the variance of the interaction estimate should be compensated by the square root of the duplication factor. The advantage of this is that the smallest bin size gets duplicated as well, decreasing the probability that any bootstrap resample leads to a divergent estimate. If the smallest bin is of size a, then the probability of not including it in a resample is e^{-a} (asymptotically). That means that the probability of at least one divergent bootstrap resample after m resamples is

$$P(\text{at least one divergent resample}) = 1 - \left(1 - e^{-a}\right)^{m} \tag{6.2}$$

Fixing m to 1000 resamples, it takes a minimum bin size of 10 to have a less than 5% chance of getting undefined resamples. At a minimum bin size of 15, this probability has become 0.03%. It thus does not seem necessary to duplicate more than 15 times.

Note, however, that this should be equivalent to constructing the asymptotic variance as is done in Section 6.2.1.a. In contract to constructing the asymptotic distribution, duplicating the data incurs a significant computational cost (linear scaling in the duplication factor) in the estimation procedure, so probably offers no clear benefit.

6.2.2 Automatically identifying the hierarchy of robust states

The dendrogram of characteristic states with the bootstrap significance values shows which branches are robust, and contains a hierarchy of states beyond those which would result from a simple distance cut-off (as was used to analyse a small dendrogram in Section 5.3.7). To automate this and extract the relevant biology, I identified states according to the following two rules:

1. *Every* robust branch is a state.

2. When Tree B is a significant branch and leaf A ∪ Tree B is also a significant branch, then this results in the states {A, B, A ∪ B }.

Note that Rule 1 implies that a single characteristic state can appear in multiple states if it is part of a robust hierarchy. Rule 2 says that if a singleton state gets robustly added to a robust tree (possible a singleton itself), then the singleton itself is also an independent state.

The states that result from these rules should give a full view on the robust states and substates present among the characteristic states of the higher-order interactions, but there are a number of issues with these rules. For example, if multiple states involve mostly the same gene expression pattern, then they tend to form a very robust branch and hierarchy. However, having these few genes result in many states might lead to redundant states. Furthermore, the branchings at very high Dice distance might be robust, but conflate different biological states. Nonetheless, these rules lead to at worst a redundant set of states that require further manual inspection.

I applied the two rules to the dendrogram of characteristic states from interactions in the merged data set of developmental neurons and astrocytes (the one used in Section 5.3.6). Cutting the dendrogram at a Dice distance of 0.88 led to 235 states, but the rules above reduced this to 185. To reveal the hierarchy implicit in these states, it makes sense to cluster this new set of states just like the previous one. However, this time it is desirable to emphasise the subset-superset relationship, which the Dice distance is not ideally suited for. Some first experiments showed that the largest states—those that includes dozens to hundreds of characteristic states—end up clustering together at low Dice-distance, rather than clearly being a superset of some of the more specialised states. A difference distance metric should therefore be found, which is planned future work. If a good distance metric can be found, then the resulting dendrogram and set of states would offer a clearer view of the robust structure present in the states, and reveal their implicit hierarchy.

6.3 Concluding remarks

In this thesis, I investigated the role maximum entropy, or Ising-like interactions in gene expression data, with an emphasis on higher-order interactions. Most current descriptions of interaction networks contain pairwise interactions only, leaving the role of higher-order dependencies in gene regulation mostly unexplored. To a certain extent, the current status quo is understandable, since higher-order interactions contradict many of our intuitions about the communication and control of complex systems in general, and gene regulation in particular. We generally think of interactions as directional, identifying a source and a target of the interaction. This direction then immediately elicits causal language—A activates B, C represses D—which shapes our thinking. In contrast, the interactions studied in this thesis are symmetric, and therefore do not always allow for a causal interpretation. Moreover, higher-order interactions cannot be easily interpreted as directional or causal at all, so are even harder to describe within the usual language of molecular biology. What language can be used to describe the regulatory effect of a triplet of genes that does not reduce it to three pairwise relationships? Still, omitting higher-order dependencies from your understanding of nature potentially hides much of the structure, and leads to underestimating the complexity of communication and control in a system. To estimate these higher-order interactions, I took two approaches: one relying on machine learning, and one on causal discovery. Both of these are active fields of research, so the challenges and results of this thesis were in part biological, and in part methodological.

I first trained restricted Boltzmann machines on various kinds of data: from simulated Ising models, to epidemiological data, to gene expression data from mouse brains. From these machines, I extracted the interactions from the model encoded in the network weights. While restricted Boltzmann machines are a particularly 'shallow' kind of neural network, I quickly ran into the kinds of problems often cited in the (deep) learning literature: interpretation, explanation, and error quantification is not easy. Interpreting and explaining the results was difficult because it is not *a priori* clear what the encoded interactions represented, and error quantification was hard because there was no natural way to obtain a measure of uncertainty for each of the estimates. This is a very general property of opaque learning systems, and one that—especially in the life sciences—has hindered their widespread adoption.

Diametrically opposed to opaque learning systems are model-free quantities that specify an estimator *purely* in terms of the data, bypassing the need to construct a model. However, this is still a philosophically slippery concept. Learning something from a truly model-free quantity would amount to inductive knowledge which, following Popper [196] and Deutsch [69], does not exist (interestingly, a special case of this is the fact that biological evolution is Darwinian and not Lamarckian). Therefore, while an individual estimator can be model-free, its interpretation cannot and draws from centuries of work on other models. Some examples of implicit assumptions present throughout this thesis are:

- The central dogma of molecular biology holds.
- Genes are a meaningful abstraction.
- Causality is a meaningful abstraction.
- The scRNA-seq protocol accurately reflects *in vivo* gene expression.
- There is functional human-mouse orthology.
- *etc.*

These are deep and complex assumptions, and more part of biological *folklore* than of any precise analysis. Nonetheless, I set out to extract knowledge about molecular mechanisms, protein interactions, cell types, and developmental lineages just by 'looking' at observational gene expression data and estimating model-free interactions among genes. In contrast to the machine learning approach, this estimator came with a natural way to quantify uncertainty in the estimates, and was interpretable directly in terms of properties of the data. One of the main features of the model-free estimator was that the interactions were conditioned on all other genes being unexpressed. This separated direct from indirect effects, which was corroborated by the independently inferred causal structure, but did not lead to stronger agreement with gold-standard networks of protein interactions relative to unconditioned correlation networks. This was somewhat surprising, but might be because a background of only unexpressed genes is biologically unrealistic. Follow-up research where the background can vary will hopefully reveal more biological interactions.

Still, the network of interactions showed important structure. In particular the higher-order interactions were present among genes with regulatory roles like transcription factors and immediate early genes. While their biochemistry is relatively well-understood, how their interactions form a complex system with regulatory and representational power is largely unknown. This thesis showed that higher-order dependencies are common, widespread, and relevant to their biological functioning, which supports the hypothesis that these genes perform logical and representational roles. A more thorough exploration of these classes of genes, and an atlas of their combinatorial regulatory rules, could be a very valuable asset to biologists. I am particularly excited about the possibility that the biochemistry of the central dogma is *universal*, *i.e.* has perhaps unbounded representational power, can be programmed, and instantiated on different substrates. This would have many practical implications for fields like synthetic biology, but also offer new theoretical insight into life itself and the source of its diversity and richness.

Beyond the mechanistic interpretation of the higher-order interactions, I used the dependencies they revealed to find structure in the population they described. This proved to be even more fruitful than the direct mechanistic interpretation. I found many biologically plausible states at a resolution beyond that which could be justified by clustering in expression space, as is currently commonly done. Interestingly, the states often implied the activity of certain biochemical pathways, like cell-cycle progression, myelin metabolism, or cell differentiation, so even though direct validation against protein-protein interactions did not yield new insights, the mechanistic function of the genes helped identify a cell type's semantics. In the interest of clarity, I separated the mechanistic research from the cell state research, but I often found myself going back and forth between these interpretations, stuck in a kind of Hegelian dialectics of gene regulation. In the end, the two interpretations synthesised a dual view on one concept: higher-order conditional dependencies in the data. Indeed, gene regulation is fundamentally implemented by biochemical mechanisms, but we understand cell states and types likewise in terms of gene expression profiles that ultimately have a biochemical cause. The two hypotheses—interactions-as-mechanism and interactions-as-states—should thus not be seen as separate and competing possibilities, but rather as two imperfect interpretations that support each other.

A large part of this thesis can be considered stamp collection or, more favourably, atlas construction. As outlined in the introductory chapter, this is of vital importance, and conducive to deeper insight into a structure as complex as biology. My hope is that this thesis provided a new page in our infinite atlas of life, one on which molecular biology is viewed through a cybernetic lens, and our intuitions about causality and networks are challenged.

Finally, I would like to reflect on the transdisciplinary nature of this research. We set out to capture a piece of molecular biology in a model from physics, but developed and encountered new ideas in maths, information theory, machine learning, and systems biology. It has been a fulfilling challenge to transgress into these new areas, and learn from the many people I had the pleasure of working with.

6.A The delta method accurately predicts the confidence intervals of genetic MFIs

In this appendix, I will briefly derive the asymptotic variance of MFIs, and show that this accurately predicts the bootstrapped variance of MFIs calculated on a data set of developmental neurons and astrocytes.

Suppose the statistic T, estimated on n samples, is asymptotically distributed normally around θ, with variance σ^2. That is, T has a standard error $\frac{\sigma}{\sqrt{n}}$, and as $n \to \infty$ its (cumulative) distribution converges to that of a normal distribution, written as

$$\sqrt{n}(T - \theta) \xrightarrow[n\to\infty]{} \mathcal{N}(0, \sigma^2) \tag{6.3}$$

Consider a differentiable transformation g of the statistic T. A Taylor approximation around θ up to first order gives

$$g(T) = g(\theta) + g'(\theta)(T - \theta) + \mathcal{O}(|T - \theta|^2) \tag{6.4}$$

Rearranging and multiplying by $\sqrt{n}$ yields

$$\sqrt{n}\,(g(T) - g(\theta)) = g'(\theta)\sqrt{n}\,(T - \theta) \tag{6.5}$$

Since T was asymptotically normal around θ, $g(T)$ is asymptotically normal around $g(\theta)$ with a variance given by $g'(\theta)^2\text{Var}(T)$. A similar analysis for a function g of a vector-valued statistic $\mathbf{T}$ with covariance matrix $\mathbf{\Sigma}$ gives the generalisation in terms of the gradient of g.

$$\text{Var}\,(g(\mathbf{T})) = (\nabla g(\mathbf{T}))^T \mathbf{\Sigma} (\nabla g(\mathbf{T})) \tag{6.6}$$

Consider an n-point MFI, or nth order log-odds ratio, as a transformation of a multivariate statistic $\mathbf{T} = (a_1, \dots, a_m, b_1, \dots, b_m)$, where $2m = 2^n$, and $\mathbf{T}$ is the list of probabilities of the different entries in a contingency table associated with a sample of joint observations:

$$\begin{aligned} I_{1\dots n} = g(\mathbf{T}) &= \log \frac{a_1 \dots a_m}{b_1 \dots b_m} &(6.7)\\ &= \sum_{i=1}^{m} \log(a_i) - \log(b_i) &(6.8)\end{aligned}$$

The gradient of g is

$$\nabla g(\mathbf{T})^T = \left(\frac{1}{a_1}, \dots, \frac{1}{a_m}, \frac{-1}{b_1}, \dots, \frac{-1}{b_m}\right) \tag{6.9}$$

The entries in a contingency table are distributed multinomially with a $2m \times 2m$ covariance matrix $\mathbf{\Sigma}$, given by

$$\mathbf{\Sigma} = n^{-1} \begin{bmatrix} a_1(1-a_1) & -a_1 a_2 & \dots & -a_1 b_m \\ -a_2 a_1 & a_2(1-a_2) & & \\ \vdots & \vdots & \ddots & \\ -b_m a_1 & -b_m a_2 & & b_m(1-b_m) \end{bmatrix} \tag{6.10}$$

It can then be easily verified that, by Equation (6.6), the variance of an MFI is proportional to the inverse harmonic mean of the probabilities and given by:

$$\mathrm{Var}\left(\log\frac{a_1\ldots a_m}{b_1\ldots b_m}\right)=n^{-1}\left(\sum_{i=1}^{m}\frac{1}{a_i}+\frac{1}{b_i}\right)\tag{6.11}$$

The univariate case of Equation (6.7) reduces to the sample logit, defined as $\mathrm{logit}(p)=\log\frac{p}{(1-p)}$. Set $a_1=P(X=1)=p$ and $b_1=P(X=0)=1-p$, then Equation (6.11) indeed retrieves the standard error on the sample logit of $\left(\sqrt{np(1-p)}\right)^{-1}$. More interestingly, the standard error $\mathrm{SE}(I_{1\ldots n})$ on an n-point MFI estimated on N samples is given by

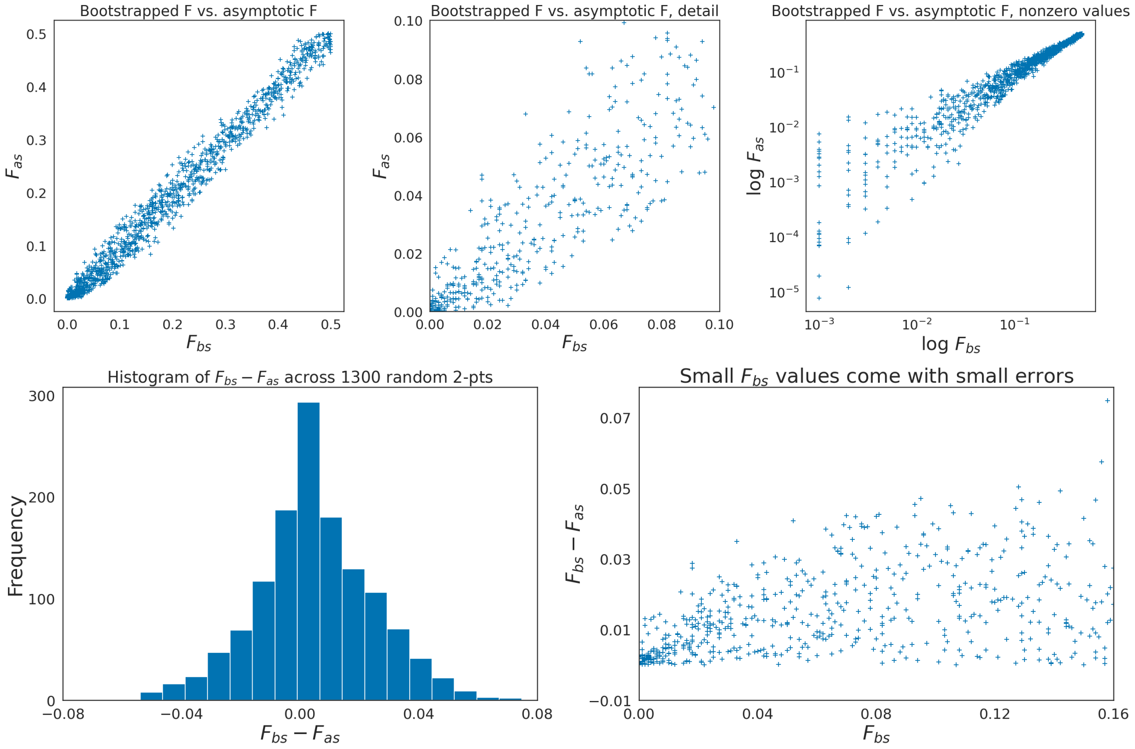
$$\mathrm{SE}(I_{1\ldots n})=\sqrt{\frac{1}{N}\sum_{i=1}^{2^n}\frac{1}{p_i}}=\sqrt{\sum_{i=1}^{2^n}\frac{1}{n_i}}\tag{6.12}$$

where p_i denotes the probability of the ith state, and n_i the ith entry from the contingency table. For example, a 2-point interaction has a standard error given by

$$\begin{aligned}\mathrm{SE}(I_{X_1,X_2})=&\left[\frac{1}{N}\left(\frac{1}{p(X_1=1,X_2=1\mid \underline{X}=0)}+\frac{1}{p(X_1=0,X_2=1\mid \underline{X}=0)}\right.\right.\\&\left.\left.+\frac{1}{p(X_1=1,X_2=0\mid \underline{X}=0)}+\frac{1}{p(X_1=0,X_2=0\mid \underline{X}=0)}\right)\right]^{1/2}\qquad(6.13)\\&=\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{01}}+\frac{1}{n_{10}}+\frac{1}{n_{00}}}\qquad(6.14)\end{aligned}$$

where n_{ab} denotes the number of samples with $X_1=a$ and $X_2=b$, conditioned on their Markov blanket being 0.

From the standard error $\mathrm{SE}(I)$, a symmetric 95% confidence interval can be constructed as $\widehat{I}\pm1.96\times\mathrm{SE}(I)$, where $\widehat{I}$ is the point estimate for the interaction. From the asymptotic variance σ^2, the F-value of an interaction I can be calculated by evaluating the cumulative distribution function of a normal distribution $\mathcal{N}(\widehat{I},\sigma^2)$ at zero. Whether the asymptotic F-values and the standard errors accurately reflect the uncertainty on finite samples depends on how quickly the transformed distribution converges to a normal distribution. To investigate this, I sampled 10,000 random pairs of genes, 1,322 of which corresponded to estimable 2-point interactions in the data set used in Figure 3.3. For each of these, I estimated the F-value with the asymptotic method outlined above (F_{as}), as well as on 1,000 bootstrap resamples (F_{bs}). Figure 6.2 shows a comparison of these two methods. It can be seen that the two estimates for the F-value are strongly correlated, and that the differences are symmetrically distributed around 0, with a maximum absolute difference of 0.075. Moreover, the difference tends to be smaller for small F-values. Of the 343 interactions that had $F_{bs}<0.05$, just ten had $F_{as}>0.05$, and none had $F_{as}>0.07$.



Bootstrapped F vs. asymptotic F

Bootstrapped F vs. asymptotic F, detail

Bootstrapped F vs. asymptotic F, nonzero values

Histogram of $F_{bs} - F_{as}$ across 1300 random 2-pts

Small F_{bs} values come with small errors

Figure 6.2: Calculating the F-value on bootstrap resamples or using the asymptotic approximation leads to mostly the same conclusions. The difference between the bootstrapped F-value and the asymptotic F-value is smaller for small F-values. From the 1,322 sampled and estimable interactions, only 5 had a difference $|F_{bs} - F_{as}| > 0.06$, the largest discrepancy being $F_{bs} - F_{as} = 0.158 - 0.083 = 0.075$.

Bibliography

[1] Dissociation of mouse embryonic neural tissue for single cell rna sequencing, 2019.

[2] User guide: Chromium single cell 3' reagent kits v2. *10X Genomics Inc.*, Manual no. CG00052 Rev F, 2019.

[3] The gene ontology resource: enriching a gold mine. *Nucleic acids research*, 49(D1):D325–D334, 2021.

[4] 10x Genomics. Transcriptional profiling of 1.3 million brain cells with the chromium single cell 3'solution. 2017.

[5] D. Adams and C. Cerf. *The Salmon of Doubt: Hitchhiking the Galaxy One Last Time*. Dirk Gently bk. 3. Harmony Books, 2002.

[6] Alan Agresti. *Categorical data analysis*. John Wiley & Sons, 2003.

[7] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, et al. Scenic: single-cell regulatory network inference and clustering. *Nature methods*, 14(11):1083–1086, 2017.

[8] Larissa Albantakis, William Marshall, Erik Hoel, and Giulio Tononi. What caused what? A quantitative account of actual causation using dynamical causal networks. aug 2017.

[9] Julieta Alfonso, Corentin Le Magueresse, Annalisa Zuccotti, Konstantin Khodosevich, and Hannah Monyer. Diazepam binding inhibitor promotes progenitor proliferation in the postnatal svz by reducing gaba signaling. *Cell stem cell*, 10(1):76–87, 2012.

[10] Joseph Altman and Gopal D Das. Autoradiographic and histological evidence of postnatal hippocampal neurogenesis in rats. *Journal of Comparative Neurology*, 124(3):319–335, 1965.

[11] Tallulah S Andrews and Martin Hemberg. M3drop: dropout-based feature selection for scrnaseq. *Bioinformatics*, 35(16):2865–2867, 2019.

[12] Yaron E Antebi, James M Linton, Heidi Klumpe, Bogdan Bintu, Mengsha Gong, Christina Su, Reed McCardell, and Michael B Elowitz. Combinatorial signal perception in the bmp pathway. *Cell*, 170(6):1184–1196, 2017.

[13] Dvir Aran, Agnieszka P Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, Ram P Naikawadi, Paul J Wolters, Adam R Abate, et al.

Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology*, 20(2):163–172, 2019.

[14] David N Arnosti, Scott Barolo, Michael Levine, and Stephen Small. The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development*, 122(1):205–214, 1996.

[15] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[16] Shahram Bahrami and Finn Drabløs. Gene regulation in the immediate-early response process. *Advances in biological regulation*, 62:37–49, 2016.

[17] Akhilesh Kumar Bajpai, Sravanthi Davuluri, Kriti Tiwary, Sithalechumi Narayanan, Sailaja Oguru, Kavyashree Basavaraju, Deena Dayalan, Kavitha Thirumurugan, and Kshitish K Acharya. Systematic comparison of the protein-protein interaction databases from a user's perspective. *Journal of Biomedical Informatics*, 103:103380, 2020.

[18] Dalia Barkley, Reuben Moncada, Maayan Pour, Deborah A Liberman, Ian Dryg, Gregor Werba, Wei Wang, Maayan Baron, Anjali Rao, Bo Xia, et al. Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. *Nature Genetics*, 54(8):1192–1201, 2022.

[19] Francesco Bartolucci, Roberto Colombi, and Antonio Forcina. An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. *Statistica Sinica*, pages 691–711, 2007.

[20] Gregory Bateson. *Steps to an Ecology of Mind*. Chandler Publishing Company, 1972.

[21] Federico Battiston, Enrico Amico, Alain Barrat, Ginestra Bianconi, Guilherme Ferraz de Arruda, Benedetta Franceschiello, Iacopo Iacopini, Sonia Kéfi, Vito Latora, Yamir Moreno, et al. The physics of higher-order interactions in complex systems. *Nature Physics*, 17(10):1093–1098, 2021.

[22] Federico Battiston, Giulia Cencetti, Iacopo Iacopini, Vito Latora, Maxime Lucas, Alice Patania, Jean Gabriel Young, and Giovanni Petri. Networks beyond pairwise interactions: Structure and dynamics, 2020.

[23] J. Baudrillard, S.F. Glaser, and University of Michigan Press. *Simulacra and Simulation*. Body, in theory. University of Michigan Press, 1994.

[24] Sjoerd Viktor Beentjes and Ava Khamseh. Higher-order interactions in statistical physics and machine learning: A model-independent solution to the inverse problem at equilibrium. *Physical Review E*, 102(5), 2020.

[25] Anthony J Bell. The co-information lattice. In *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation: ICA*, volume 2003. Citeseer, 2003.

[26] Nitin Bhardwaj and Hui Lu. Correlation between gene expression profiles and protein–protein interactions within and across genomes. *Bioinformatics*, 21(11):2730–2738, 2005.

[27] Yaqiong Bi, Song Chen, Jiazhi Jiang, Jie Yao, Gang Wang, Qiang Zhou, and Sheng Li. Cdca8 expression and its clinical relevance in patients with bladder cancer. *Medicine*, 97(34), 2018.

[28] Marta Biagioli, Milena Pinto, Daniela Cesselli, Marta Zaninello, Dejan Lazarevic, Paola Roncaglia, Roberto Simone, Christina Vlachouli, Charles Plessy, Nicolas Bertin, et al. Unexpected expression of α-and β-globin in mesencephalic dopaminergic neurons and glial cells. *Proceedings of the National Academy of Sciences*, 106(36):15454–15459, 2009.

[29] Christoph Biesemann, Mads Grønborg, Elisa Luquet, Sven P Wichert, Véronique Bernard, Simon R Bungers, Ben Cooper, Frédérique Varoqueaux, Liyi Li, Jennifer A Byrne, et al. Proteomic screening of glutamatergic mouse brain synaptosomes isolated by fluorescence activated sorting. *The EMBO journal*, 33(2):157–170, 2014.

[30] Christopher S Bjornsson, Maria Apostolopoulou, Yangzi Tian, and Sally Temple. It takes a village: constructing the neurogenic niche. *Developmental cell*, 32(4):435–446, 2015.

[31] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[32] Jeroen FJ Bogie, Elien Grajchen, Elien Wouters, Aida Garcia Corrales, Tess Dierckx, Sam Vanherle, Jo Mailleux, Pascal Gervois, Esther Wolfs, Jonas Dehairs, et al. Stearoyl-coa desaturase-1 impairs the reparative properties of macrophages and microglia in the brain. *Journal of Experimental Medicine*, 217(5), 2020.

[33] Ann Boija, Isaac A Klein, Benjamin R Sabari, Alessandra Dall'Agnese, Eliot L Coffey, Alicia V Zamudio, Charles H Li, Krishna Shrinivas, John C Manteiga, Nancy M Hannett, et al. Transcription factors activate genes through the phase-separation capacity of their activation domains. *Cell*, 175(7):1842–1855, 2018.

[34] Dafna Bonneh-Barkay, Guoji Wang, Adam Starkey, Ronald L Hamilton, and Clayton A Wiley. In vivo chi3l1 (ykl-40) expression in astrocytes in acute and chronic neurological diseases. *Journal of neuroinflammation*, 7(1):1–8, 2010.

[35] Michael J Borrett, Brendan T Innes, Nareh Tahmasian, Gary D Bader, David R Kaplan, and Freda D Miller. A shared transcriptional identity for forebrain and dentate gyrus neural stem cells from embryogenesis to adulthood. *Eneuro*, 9(1), 2022.

[36] Linda L Boshans, Heun Soh, William M Wood, Timothy M Nolan, Ion I Mandoiu, Yuchio Yanagawa, Anastasios V Tzingounis, and Akiko Nishiyama. Direct reprogramming of oligodendrocyte precursor cells into gabaergic inhibitory neurons by a single homeodomain transcription factor dlx2. *Scientific reports*, 11(1):1–15, 2021.

[37] Gerard A Bouland, Ahmed Mahfouz, and Marcel JT Reinders. Differential analysis of binarized single-cell rna sequencing data captures biological variation. *NAR genomics and bioinformatics*, 3(4):lqab118, 2021.

[38] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7):1177–1186, 2017.

[39] Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C Marioni, et al. Accounting for technical noise in single-cell rna-seq experiments. *Nature methods*, 10(11):1093–1095, 2013.

[40] J. Brockman. *Possible Minds: Twenty-Five Ways of Looking at AI*. Penguin Press, 2019.

[41] Jelle Bruineberg, Krzysztof Dolega, Joe Dewhurst, and Manuel Baltieri. The emperor's new markov blankets. *Behavioral and Brain Sciences*, pages 1–63, 2020.

[42] William S Burroughs. *Naked Lunch*. Grove, 1959.

[43] John D Cahoy, Ben Emery, Amit Kaushal, Lynette C Foo, Jennifer L Zamanian, Karen S Christopherson, Yi Xing, Jane L Lubischer, Paul A Krieg, Sergey A Krupenko, et al. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *Journal of Neuroscience*, 28(1):264–278, 2008.

[44] Evan P Campbell, Ian K Quigley, and Chris Kintner. Foxn4 promotes gene expression required for the formation of multiple motile cilia. *Development*, 143(24):4654–4664, 2016.

[45] Meritxell Canals, Daniel Marcellino, Francesca Fanelli, Francisco Ciruela, Piero De Benedetti, Steven R Goldberg, Kim Neve, Kjell Fuxe, Luigi F Agnati, Amina S Woods, et al. Adenosine a2a-dopamine d2 receptor-receptor heteromerization: qualitative and quantitative assessment by fluorescence and bioluminescence energy transfer. *Journal of Biological Chemistry*, 278(47):46741–46749, 2003.

[46] George T Cantwell, Yanchen Liu, Benjamin F Maier, Alice C Schwarze, Carlos A Serván, Jordan Snyder, and Guillaume St-Onge. Thresholding normally distributed data creates complex networks. *Physical Review E*, 101(6):062302, 2020.

[47] Junyue Cao, Jonathan S Packer, Vijay Ramani, Darren A Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N Furlan, Frank J Steemers, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667, 2017.

[48] Sofie Carmans, Jerome JA Hendriks, Kristof Thewissen, Jimmy Van den Eynden, Piet Stinissen, Jean-Michel Rigo, and Niels Hellings. The inhibitory neurotransmitter glycine modulates macrophage activity by activation of neutral amino acid transporters. *Journal of neuroscience research*, 88(11):2420–2430, 2010.

[49] Marjolein MJ Caron, Maxime Eveque, Berta Cillero-Pastor, Ron Heeren, Bas Housmans, Kasper Derks, Andy Cremers, Mandy J Peffers, Lodewijk W van Rhijn, Guus

van den Akker, et al. Sox9 determines translational capacity during early chondrogenic differentiation of atdc5 cells by regulating expression of ribosome biogenesis factors and ribosomal proteins. *Frontiers in cell and developmental biology*, 9:1489, 2021.

[50] Miguel A Carreira-Perpinan and Geoffrey Hinton. On contrastive divergence learning. In *International workshop on artificial intelligence and statistics*, pages 33–40. PMLR, 2005.

[51] Tara Chari, Brandon Weissbourd, Jase Gehring, Anna Ferraioli, Lucas Leclère, Makenna Herl, Fan Gao, Sandra Chevalier, Richard Copley, Evelyn Houliston, et al. Whole animal multiplexed single-cell rna-seq reveals plasticity of clytia medusa cell types. 2021.

[52] Meng Chen, Till B Puschmann, Pavel Marasek, Masaki Inagaki, Marcela Pekna, Ulrika Wilhelmsson, and Milos Pekny. Increased neuronal differentiation of neural progenitor cells derived from phosphovimentin-deficient mice. *Molecular Neurobiology*, 55(7):5478–5489, 2018.

[53] Yan Chen, Lijun Zhu, Zhengmei Fang, Yuelong Jin, Chong Shen, Yingshui Yao, and Chengchao Zhou. Soluble guanylate cyclase contribute genetic susceptibility to essential hypertension in the han chinese population. *Annals of Translational Medicine*, 7(22), 2019.

[54] Won-Ki Cho, Jan-Hendrik Spille, Micca Hecht, Choongman Lee, Charles Li, Valentin Grube, and Ibrahim I Cisse. Mediator and rna polymerase ii clusters associate in transcription-dependent condensates. *Science*, 361(6400):412–415, 2018.

[55] Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 2015.

[56] Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic acids research*, 47(D1):D330–D338, 2019.

[57] Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562(7727):367–372, 2018.

[58] Tabula Sapiens Consortium*, Robert C Jones, Jim Karkanias, Mark A Krasnow, Angela Oliveira Pisco, Stephen R Quake, Julia Salzman, Nir Yosef, Bryan Bulthaup, Phillip Brown, et al. The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896, 2022.

[59] Guido Cossu, Luigi Del Debbio, Tommaso Giani, Ava Khamseh, and Michael Wilson. Machine learning determination of dynamical parameters: The Ising model case. *Physical Review B*, 2019.

[60] Francis HC Crick. On protein synthesis. In *Symp Soc Exp Biol*, volume 12, page 8, 1958.

[61] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[62] Xiao-Han Cui, Qiu-Ju Peng, Ren-Zhi Li, Xia-Jie Lyu, Chun-Fu Zhu, and Xi-Hu Qin. Cell division cycle associated 8: A novel diagnostic and prognostic biomarker for hepatocellular carcinoma. *Journal of cellular and molecular medicine*, 25(24):11097–11112, 2021.

[63] Charles Darwin. *On the origin of species*. London, Murray, 1859.

[64] Alberto De La Fuente, Nan Bing, Ina Hoeschele, and Pedro Mendes. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574, 2004.

[65] P.S. de Laplace. *Théorie analytique des probabilités*. Paris, 1814.

[66] Eulàlia De Nadal, Gustav Ammerer, and Francesc Posas. Controlling gene expression in response to stress. *Nature Reviews Genetics*, 12(12):833–845, 2011.

[67] Conor Delaney, Alexandra Schnell, Louis V Cammarata, Aaron Yao-Smith, Aviv Regev, Vijay K Kuchroo, and Meromit Singer. Combinatorial prediction of marker panels from single-cell transcriptomic data. *Molecular systems biology*, 15(10):e9005, 2019.

[68] Guillaume Desjardins, Aaron Courville, and Yoshua Bengio. Adaptive parallel tempering for stochastic maximum likelihood learning of rbms. *arXiv preprint arXiv:1012.3476*, 2010.

[69] David Deutsch. *The beginning of infinity: Explanations that transform the world*. penguin uK, 2011.

[70] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature biotechnology*, 35(4):316–319, 2017.

[71] Florian Duclot and Mohamed Kabbaj. The role of early growth response 1 (egr1) in brain plasticity and neuropsychiatric disorders. *Frontiers in behavioral neuroscience*, 11:35, 2017.

[72] Bianca Dumitrascu, Soledad Villar, Dustin G Mixon, and Barbara E Engelhardt. Optimal marker gene selection for cell type discrimination in single cell analyses. *Nature communications*, 12(1):1–8, 2021.

[73] Bradley Efron, Elizabeth Halloran, and Susan Holmes. Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences*, 93(23):13429–13429, 1996.

[74] Marc G Elgort, John M O'Shea, Yike Jiang, and Donald E Ayer. Transcriptional and translational downregulation of thioredoxin interacting protein is required for metabolic reprogramming during g1. *Genes & cancer*, 1(9):893–907, 2010.

[75] Frank Emmert-Streib, Galina Glazko, Ricardo De Matos Simoes, et al. Statistical inference and reverse engineering of gene regulatory networks from observational expression data. *Frontiers in genetics*, 3:8, 2012.

[76] Kurt Engeland. Cell cycle arrest through indirect transcriptional repression by p53: I have a dream. *Cell Death & Differentiation*, 25(1):114–132, 2018.

[77] Amy E Evans, CM Kelly, Sophie Victoria Precious, and Anne Elizabeth Rosser. Molecular regulation of striatal development: a review. *Anatomy research international*, 2012, 2012.

[78] Myron K Evans, Yurika Matsui, Beisi Xu, Catherine Willis, Jennifer Loome, Luis Milburn, Yiping Fan, Vishwajeeth Pagala, and Jamy C Peng. Ybx1 fine-tunes prc2 activities to control embryonic brain development. *Nature communications*, 11(1):1–18, 2020.

[79] Jeremiah J Faith, Boris Hayete, Joshua T Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J Collins, and Timothy S Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, 5(1):e8, 2007.

[80] Sacri R Ferrón, Marika Charalambous, Elizabeth Radford, Kirsten McEwen, Hendrik Wildner, Eleanor Hind, Jose Manuel Morante-Redolat, Jorge Laborda, Francois Guillemot, Steven R Bauer, et al. Postnatal loss of dlk1 imprinting in stem cells and niche astrocytes regulates neurogenesis. *Nature*, 475(7356):381–385, 2011.

[81] Asja Fischer. Training restricted boltzmann machines. *KI-Künstliche Intelligenz*, 29(4):441–444, 2015.

[82] Asja Fischer and Christian Igel. Empirical analysis of the divergence of Gibbs sampling based learning algorithms for restricted Boltzmann machines. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010.

[83] Patrick Forré and Joris M Mooij. Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. *arXiv preprint arXiv:1807.03024*, 2018.

[84] Yoav Freund and David Haussler. Unsupervised learning of distributions of binary vectors using two layer networks. 1994.

[85] Nir Friedman and Daphne Koller. Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks. *Machine learning*, 50(1):95–125, 2003.

[86] David J. Galas, James Kunert-Graf, Lisa Uechi, and Nikita A. Sakhanenko. Towards an information theory of quantitative genetics, 2019.

[87] David J Galas, Nikita A Sakhanenko, Alexander Skupin, and Tomasz Ignac. Describing the complexity of systems: Multivariable “set complexity” and the information basis of systems biology. *Journal of Computational Biology*, 21(2):118–140, 2014.

[88] Lucy L Gao, Jacob Bien, and Daniela Witten. Selective inference for hierarchical clustering. *arXiv preprint arXiv:2012.02936*, 2020.

[89] José Garcıa-Martınez, Agustın Aranda, and José E Pérez-Ortın. Genomic run-on evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms. *Molecular cell*, 15(2):303–313, 2004.

[90] Sonia Garel, Faustino Marín, Rudolf Grosschedl, and Patrick Charnay. Ebf1 controls early cell differentiation in the embryonic striatum. *Development*, 126(23):5285–5294, 1999.

[91] Ankur Garg, Abdul Hannan, Qian Wang, Neoklis Makrides, Jian Zhong, Hongge Li, Sungtae Yoon, Yingyu Mao, and Xin Zhang. Etv transcription factors functionally diverge from their upstream fgf signaling in lens development. *Elife*, 9:e51915, 2020.

[92] Daniel M Gatti, William T Barry, Andrew B Nobel, Ivan Rusyn, and Fred A Wright. Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC genomics*, 11(1):1–10, 2010.

[93] Shila Ghazanfar, Yingxin Lin, Xianbin Su, David Ming Lin, Ellis Patrick, Ze Guang Han, John C. Marioni, and Jean Yee Hwa Yang. Investigating higher-order interactions in single-cell data with scHOT. *Nature Methods*, 2020.

[94] Jesse Gillis, Sara Ballouz, and Paul Pavlidis. Bias tradeoffs in the creation and analysis of protein–protein interaction networks. *Journal of proteomics*, 100:44–54, 2014.

[95] Ohad Givaty and Yaakov Levy. Protein sliding along dna: dynamics and structural characterization. *Journal of molecular biology*, 385(4):1087–1097, 2009.

[96] Garique FV Glonek and Peter McCullagh. Multivariate logistic models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(3):533–546, 1995.

[97] Magdalena Götz and Wieland B Huttner. The cell biology of neurogenesis. *Nature reviews Molecular cell biology*, 6(10):777–788, 2005.

[98] Alexander Grothendieck and Roy Lisker. *Récoltes et Semailles*. Gallimard, 1986.

[99] Aran Groves, Yasuyuki Kihara, Deepa Jonnalagadda, Richard Rivera, Grace Kennedy, Mark Mayford, and Jerold Chun. A functionally defined in vivo astrocyte population identified by c-fos activation in a mouse model of multiple sclerosis modulated by s1p signaling: immediate-early astrocytes (ieastrocytes). *ENeuro*, 5(5), 2018.

[100] Hongshan Guo, Gabriel Golczer, Ben S Wittner, Adam Langenbucher, Marcus Zachariah, Taronish D Dubash, Xin Hong, Valentine Comaills, Risa Burr, Richard Y Ebright, et al. Nr4a1 regulates expression of immediate early genes, suppressing replication stress in cancer. *Molecular Cell*, 81(19):4041–4058, 2021.

[101] Aaron J Gutknecht, Michael Wibral, and Abdullah Makkeh. Bits and pieces: Understanding information decomposition from part-whole relationships and formal logic. *Proceedings of the Royal Society A*, 477(2251):20210110, 2021.

[102] Ryuji Hamamoto, Yoichi Furukawa, Masashi Morita, Yuko Iimura, Fabio Pittella Silva, Meihua Li, Ryuichiro Yagyu, and Yusuke Nakamura. Smyd3 encodes a histone methyltransferase involved in the proliferation of cancer cells. *Nature cell biology*, 6(8):731–740, 2004.

[103] Matthew C Havrda, Brent T Harris, Akio Mantani, Nora M Ward, Brenton R Paolella, Verginia C Cuzon, Hermes H Yeh, and Mark A Israel. Id2 is required for specification of dopaminergic neurons during adult olfactory neurogenesis. *Journal of Neuroscience*, 28(52):14074–14087, 2008.

[104] Jing He, Michael Kleyman, Jianjiao Chen, Aydin Alikaya, Kathryn M Rothenhoefer, Bilge Esin Ozturk, Morgan Wirthlin, Andreea C Bostan, Kenneth Fish, Leah C Byrne, et al. Transcriptional and anatomical diversity of medium spiny neurons in the primate striatum. *Current Biology*, 31(24):5473–5486, 2021.

[105] David Heckerman and Dan Geiger. Learning bayesian networks: a unification for discrete and gaussian domains. *arXiv preprint arXiv:1302.4957*, 2013.

[106] KH Herzog and JI Morgan. Cellular immediate-early genes and cell death in the nervous system. *Neuropathology and Applied Neurobiology*, 22(6):484–488, 1996.

[107] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 2002.

[108] Geoffrey E Hinton. Boltzmann machine. *Scholarpedia*, 2(5):1668, 2007.

[109] Hannah Hochgerner, Amit Zeisel, Peter Lönnerberg, and Sten Linnarsson. Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell rna sequencing. *Nature neuroscience*, 21(2):290–299, 2018.

[110] Erik Hoel and Michael Levin. Emergence of informative higher scales in biological systems: a computational toolkit for optimal prediction and control. *Communicative and Integrative Biology*, 2020.

[111] Douglas R Hofstadter. *I am a strange loop*. Basic books, 2007.

[112] Frantisek Honti, Stephen Meader, and Caleb Webber. Unbiased Functional Clustering of Gene Variants with a Phenotypic-Linkage Network. *PLoS Computational Biology*, 2014.

[113] Tycho M Hoogland, Bernd Kuhn, Werner Göbel, Wenying Huang, Junichi Nakai, Fritjof Helmchen, Jane Flint, and Samuel S-H Wang. Radially expanding transglial calcium waves in the intact cerebellum. *Proceedings of the National Academy of Sciences*, 106(9):3496–3501, 2009.

[114] Bidossessi Wilfried Hounkpe, Francine Chenou, Franciele de Lima, and Erich Vinicius De Paula. Hrt atlas v1. 0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive rna-seq datasets. *Nucleic acids research*, 49(D1):D947–D955, 2021.

[115] Chiaowen Joyce Hsiao, PoYuan Tung, John D Blischak, Jonathan E Burnett, Kenneth A Barr, Kushal K Dey, Matthew Stephens, and Yoav Gilad. Characterizing and inferring quantitative cell cycle phase in single-cell rna-seq data analysis. *Genome research*, 30(4):611–621, 2020.

[116] Hui Hu, Ya-Ru Miao, Long-Hao Jia, Qing-Yang Yu, Qiong Zhang, and An-Yuan Guo. Animaltfdb 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic acids research*, 47(D1):D33–D38, 2019.

[117] Hsi-Yuan Huang, Yang-Chi-Dung Lin, Jing Li, Kai-Yao Huang, Sirjana Shrestha, Hsiao-Chin Hong, Yun Tang, Yi-Gang Chen, Chen-Nan Jin, Yuan Yu, et al. mirtarbase 2020: updates to the experimentally validated microrna–target interaction database. *Nucleic acids research*, 48(D1):D148–D154, 2020.

[118] Curtis Huttenhower, Matthew A Hibbs, Chad L Myers, Amy A Caudy, David C Hess, and Olga G Troyanskaya. The impact of incomplete knowledge on evaluation: an experimental benchmark for protein function prediction. *Bioinformatics*, 25(18):2404–2410, 2009.

[119] Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):e12776, 2010.

[120] Eva Jablonka, Michael Lachmann, and Marion J Lamb. Evidence, mechanisms and models for the inheritance of acquired characters. *Journal of theoretical biology*, 158(2):245–268, 1992.

[121] Allyn Jackson. Comme appelé du néant - as if summoned from the void: The life of alexandre grothendieck. *Notices of the Amer. Math. Soc.*, 51:1038 – 1056, 2004.

[122] F. Jacob. *La logique du vivant: une histoire de l'hérédité*. Bibliothèque des sciences humaines. Gallimard, 1970.

[123] Katherine James, SJ Lycett, A Wipat, and JS Hallinan. Multiple gold standards address bias in functional network integration. *School of Computing Science Technical Report Series*, 2011.

[124] Ryan G James and James P Crutchfield. Multivariate dependence beyond shannon information. *Entropy*, 19(10):531, 2017.

[125] Abel Jansma. Higher-order in-and-outeractions reveal synergy and logical dependence beyond shannon-information. *arXiv preprint arXiv:2205.04440*, 2022.

[126] Abel Jansma. Higher-order interactions and their duals reveal synergy and logical dependence beyond shannon-information. *Entropy*, 25(4):648, 2023.

[127] E. T. Jaynes. Information theory and statistical mechanics. II. *Physical Review*, 1957.

[128] Arttu Jolma, Yimeng Yin, Kazuhiro R. Nitta, Kashyap Dave, Alexander Popov, Minna Taipale, Martin Enge, Teemu Kivioja, Ekaterina Morgunova, and Jussi Taipale. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, 2015.

[129] Joanna Kaczynski, Tiffany Cook, and Raul Urrutia. Sp1-and krüppel-like transcription factors. *Genome biology*, 4(2):1–8, 2003.

[130] Hyunju Kang, Jejoong Yoo, Byeong-Kwon Sohn, Seung-Won Lee, Hong Soo Lee, Wenjie Ma, Jung-Min Kee, Aleksei Aksimentiev, and Hajin Kim. Sequence-dependent dna condensation as a driving force of dna phase separation. *Nucleic acids research*, 46(18):9401–9413, 2018.

[131] Stuart A Kauffman et al. *The origins of order: Self-organization and selection in evolution*. Oxford University Press, USA, 1993.

[132] Samuel Kerrien, Sandra Orchard, Luisa Montecchi-Palazzi, Bruno Aranda, Antony F Quinn, Nisha Vinod, Gary D Bader, Ioannis Xenarios, Jérôme Wojcik, David Sherman, et al. Broadening the horizon–level 2.5 of the hupo-psi format for molecular interactions. *BMC biology*, 5(1):1–11, 2007.

[133] Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740–742, 2014.

[134] Sangjin Kim, Erik Broströmer, Dong Xing, Jianshi Jin, Shasha Chong, Hao Ge, Siyuan Wang, Chan Gu, Lijiang Yang, Yi Qin Gao, et al. Probing allostery through dna. *Science*, 339(6121):816–819, 2013.

[135] Vladimir Yu Kiselev, Andrew Yiu, and Martin Hemberg. scmap: projection of single-cell rna-seq data across data sets. *Nature methods*, 15(5):359–362, 2018.

[136] Nathalie Kliemann, Neil Murphy, Vivian Viallon, Heinz Freisling, Konstantinos K Tsilidis, Sabina Rinaldi, Francesca R Mancini, Guy Fagherazzi, Marie-Christine Boutron-Ruault, Heiner Boeing, et al. Predicted basal metabolic rate and cancer risk in the european prospective investigation into cancer and nutrition. *International journal of cancer*, 147(3):648–661, 2020.

[137] Heidi E Klumpe. *Context-dependent, combinatorial logic of BMP signaling*. PhD thesis, California Institute of Technology, 2021.

[138] Teng Wei Koay, Carina Osterhof, Ilaria MC Orlando, Anna Keppner, Daniel Andre, Schayan Yousefian, María Suárez Alonso, Miguel Correia, Robert Markworth, Johannes Schödel, et al. Androglobin gene expression patterns and foxj1-dependent regulation indicate its functional association with ciliogenesis. *Journal of Biological Chemistry*, 296, 2021.

[139] Arnold Kriegstein and Arturo Alvarez-Buylla. The glial nature of embryonic and adult neural stem cells. *Annual review of neuroscience*, 32:149, 2009.

[140] Jan Krumsiek, Karsten Suhre, Thomas Illig, Jerzy Adamski, and Fabian J Theis. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC systems biology*, 5(1):21, 2011.

[141] Jack Kuipers, Polina Suter, and Giusi Moffa. Efficient Structure Learning and Sampling of Bayesian Networks. pages 1–40, 2018.

[142] Elena Kuzmin, Benjamin VanderSluis, Wen Wang, Guihong Tan, Raamesh Deshpande, Yiqun Chen, Matej Usaj, Attila Balint, Mojca Mattiazzi Usaj, Jolanda Van Leeuwen, Elizabeth N. Koch, Carles Pons, Andrius J. Dagilis, Michael Pryszlak, Jason Zi Yang Wang, Julia Hanchard, Margot Riggi, Kaicong Xu, Hamed Heydari, Bryan Joseph San Luis, Ermira Shuteriqi, Hongwei Zhu, Nydia Van Dyk, Sara Sharifpoor, Michael Costanzo, Robbie Loewith, Amy Caudy, Daniel Bolnick, Grant W. Brown, Brenda J. Andrews, Charles Boone, and Chad L. Myers. Systematic analysis of complex genetic interactions. *Science*, 2018.

[143] Gioele La Manno, Kimberly Siletti, Alessandro Furlan, Daniel Gyllborg, Elin Vinsland, Alejandro Mossi Albiach, Christoffer Mattsson Langseth, Irina Khven, Alex R Lederer, Lisa M Dratva, et al. Molecular architecture of the developing mouse brain. *Nature*, 596(7870):92–96, 2021.

[144] Raphael Lamprecht and Yadin Dudai. Differential modulation of brain immediate early genes by intraperitoneal licl. *Neuroreport: An International Journal for the Rapid Communication of Research in Neuroscience*, 1995.

[145] Caleb A Lareau, Fabiana M Duarte, Jennifer G Chew, Vinay K Kartha, Zach D Burkett, Andrew S Kohlway, Dmitry Pokholok, Martin J Aryee, Frank J Steemers, Ronald Lebofsky, et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nature Biotechnology*, 37(8):916–924, 2019.

[146] Thuc Duy Le, Tao Hoang, Jiuyong Li, Lin Liu, Huawen Liu, and Shu Hu. A fast PC algorithm for high dimensional causal discovery with multi-core PCs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019.

[147] Ed S Lein, Michael J Hawrylycz, Nancy Ao, Mikael Ayres, Amy Bensinger, Amy Bernard, Andrew F Boe, Mark S Boguski, Kevin S Brockway, Emi J Byrnes, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445(7124):168–176, 2007.

[148] Timothy R. Lezon, Jayanth R. Banavar, Marek Cieplak, Amos Maritan, and Nina V. Fedoroff. Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 2006.

[149] Ruoxin Li and Gerald Quon. ScBFA: Modeling detection patterns to mitigate technical noise in large-scale single-cell genomics data. *Genome Biology*, 2019.

[150] Yao-Yun Liang, F Charles Brunicardi, and Xia Lin. Smad3 mediates immediate early induction of id1 by tgf-β. *Cell research*, 19(1):140–148, 2009.

[151] Jiancheng Liu, Xiwei Wu, and Qiang Lu. Molecular divergence of mammalian astrocyte progenitor cells at early gliogenesis. *Development*, 149(5):dev199985, 2022.

[152] Jason W Locasale and Alejandro Wolf-Yadlin. Maximum entropy reconstructions of dynamic signaling networks from quantitative proteomics data. *PloS one*, 4(8):e6522, 2009.

[153] Elizabeth K Lucas, Sarah E Dougherty, Laura J McMeekin, Alisa T Trinh, Courtney S Reid, and Rita M Cowell. Developmental alterations in motor coordination and medium spiny neuron markers in mice lacking pgc-1α. 2012.

[154] Ward Lutz, Elena M Frank, Theodore A Craig, Richele Thompson, Ronald A Venters, Doug Kojetin, John Cavanagh, and Rajiv Kumar. Calbindin d28k interacts with ran-binding protein m: identification of interacting domains by nmr spectroscopy. *Biochemical and biophysical research communications*, 303(4):1186–1192, 2003.

[155] FeiFei Ma, Cheng Zhi, Minling Wang, Tao Li, Shahzad Akbar Khan, Zhaoen Ma, Zhiliang Jing, Chen Bo, Qiang Zhou, Shaomei Xia, et al. Dysregulated nf-κb signal promotes the hub gene pclaf expression to facilitate nasopharyngeal carcinoma proliferation and metastasis. *Biomedicine & Pharmacotherapy*, 125:109905, 2020.

[156] Tobias Maier, Marc Güell, and Luis Serrano. Correlation of mrna and protein in complex biological samples. *FEBS letters*, 583(24):3966–3973, 2009.

[157] Paolo Malatesta, Irene Appolloni, and Filippo Calzolari. Radial glia and neural stem cells. *Cell and tissue research*, 331(1):165–178, 2008.

[158] Paolo Malatesta, Eva Hartfuss, and M Gotz. Isolation of radial glial cells by fluorescent-activated cell sorting reveals a neuronal lineage. *Development*, 127(24):5253–5263, 2000.

[159] Adam A. Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 2006.

[160] Verónica Martínez-Cerdeño and Stephen C Noctor. Neural progenitor cell terminology. *Frontiers in neuroanatomy*, 12:104, 2018.

[161] HR Maturana and FJ Varela. *Autopoiesis and cognition: The realization of the living*. D. Reidel Publishing Company, Dordrecht, Holland, 1980.

[162] Gregor Mendel. *Versuche über Pflanzenhybriden*. Harvard University Press, 1965.

[163] Lina Merchan and Ilya Nemenman. On the Sufficiency of Pairwise Interactions in Maximum Entropy Models of Networks. *Journal of Statistical Physics*, 2016.

[164] Florian T Merkle, Anthony D Tramontin, José Manuel García-Verdugo, and Arturo Alvarez-Buylla. Radial glia give rise to adult neural stem cells in the subventricular zone. *Proceedings of the National Academy of Sciences*, 101(50):17528–17532, 2004.

[165] Simone Mesman, Reinier Bakker, and Marten P Smidt. Tcf4 is required for correct brain development during embryogenesis. *Molecular and cellular neuroscience*, 106:103502, 2020.

[166] Monica R Metea and Eric A Newman. Calcium signaling in specialized glial cells. *Glia*, 54(7):650–655, 2006.

[167] Tarjei Mikkelsen. Chromium single cell solutions, 2016.

[168] Prasad Minakshi, Rajesh Kumar, Mayukh Ghosh, Hari Mohan Saini, Koushlesh Ranjan, Basanti Brar, and Gaya Prasad. Single-cell proteomics: technology and applications. In *Single-Cell Omics*, pages 283–318. Elsevier, 2019.

[169] Guido Montufar and Nihat Ay. Refinements of universal approximation results for deep belief networks and restricted boltzmann machines. *Neural computation*, 23(5):1306–1319, 2011.

[170] Thierry Mora and William Bialek. Are Biological Systems Poised at Criticality?, 2011.

[171] Samantha A Morris. The evolving concept of cell identity in the single cell era. *Development*, 146(12):dev169748, 2019.

[172] Lisa Muniz, Estelle Nicolas, and Didier Trouche. Rna polymerase ii speed: a key player in controlling and adapting transcriptome composition. *The EMBO Journal*, 40(15):e105740, 2021.

[173] So Nakagawa, Stephen S Gisselbrecht, Julia M Rogers, Daniel L Hartl, and Martha L Bulyk. Dna-binding specificity changes in the evolution of forkhead transcription factors. *Proceedings of the National Academy of Sciences*, 110(30):12349–12354, 2013.

[174] Elly Nedivi, Dana Hevroni, Dorit Naot, David Israeli, and Yoav Citri. Numerous candidate plasticity-related genes revealed by differential cdna cloning. *Nature*, 363(6431):718–722, 1993.

[175] Cyril Neftel, Julie Laffy, Mariella G Filbin, Toshiro Hara, Marni E Shore, Gilbert J Rahme, Alyssa R Richman, Dana Silverbush, McKenzie L Shaw, Christine M Hebert, et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell*, 178(4):835–849, 2019.

[176] Ilya Nemenman. Information theory, multivariate dependence, and genetic network inference. *arXiv preprint q-bio/0406015*, 2004.

[177] I. Newton. *The Mathematical Principles of Natural Philosophy*. Philosophiae Naturalis Principia Mathematica. 1729.

[178] H. Chau Nguyen, Riccardo Zecchina, and Johannes Berg. Inverse statistical problems: from the inverse Ising problem to data science. *Advances in Physics*, 2017.

[179] Stephen C Noctor, Alexander C Flint, Tamily A Weissman, Ryan S Dammerman, and Arnold R Kriegstein. Neurons derived from radial glial cells establish radial units in neocortex. *Nature*, 409(6821):714–720, 2001.

[180] Stephen C Noctor, Alexander C Flint, Tamily A Weissman, Winston S Wong, Brian K Clinton, and Arnold R Kriegstein. Dividing precursor cells of the embryonic cortical ventricular zone have morphological and molecular characteristics of radial glia. *Journal of Neuroscience*, 22(8):3161–3173, 2002.

[181] Stephen C Noctor, Verónica Martínez-Cerdeño, Lidija Ivic, and Arnold R Kriegstein. Cortical neurons arise in symmetric and asymmetric division zones and migrate through specific phases. *Nature neuroscience*, 7(2):136–144, 2004.

[182] Souichi Ogata, Junji Morokuma, Tadayoshi Hayata, Gabriel Kolle, Christof Niehrs, Naoto Ueno, and Ken WY Cho. Tgf-β signaling-mediated morphogenesis: modulation of cell adhesion via cadherin endocytosis. *Genes & development*, 21(14):1817–1831, 2007.

[183] M. Oliver. *Evidence: Poems*. Beacon Press, 2010.

[184] Vikram R Paralkar, Cristian C Taborda, Peng Huang, Yu Yao, Andrew V Kossenkov, Rishi Prasad, Jing Luan, James OJ Davies, Jim R Hughes, Ross C Hardison, et al. Unlinking an lncrna from its associated cis element. *Molecular cell*, 62(1):104–110, 2016.

[185] Javier Pardo-Diaz, Lyuba V Bozhilova, Mariano Beguerisse-Diaz, Philip S Poole, Charlotte M Deane, and Gesine Reinert. Robust gene coexpression networks using signed distance correlation. *bioRxiv*, page 2020.06.21.163543, 2020.

[186] Elissa D Pastuzyn, Cameron E Day, Rachel B Kearns, Madeleine Kyrke-Smith, Andrew V Taibi, John McCormick, Nathan Yoder, David M Belnap, Simon Erlendsson, Dustin R Morado, et al. The neuronal gene arc encodes a repurposed retrotransposon gag protein that mediates intercellular rna transfer. *Cell*, 172(1-2):275–288, 2018.

[187] Ausvydas Patasius, Vincas Urbonas, and Giedre Smailyte. Skin melanoma and subsequent risk of prostate cancer: A lithuanian cancer registry study. *International journal of environmental research and public health*, 16(20):3915, 2019.

[188] Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, et al. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014.

[189] J. Pearl and D. Mackenzie. *The Book of why: The New Science of Cause and Effect*. An Allen Lane book. Penguin Books, 2019.

[190] Judea Pearl et al. Causality: Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19:2, 2000.

[191] Vini Pereira, David Waxman, and Adam Eyre-Walker. A problem with the correlation coefficient as a measure of gene expression divergence. *Genetics*, 183(4):1597–1600, 2009.

[192] Magdalena A Petryniak, Gregory B Potter, David H Rowitch, and John LR Rubenstein. Dlx1 and dlx2 control neuronal versus oligodendroglial cell fate acquisition in the developing forebrain. *Neuron*, 55(3):417–433, 2007.

[193] Pablo Picasso and William Fifield. Pablo picasso: A composite interview. *The Paris Review 32*, 1964.

[194] Plato. *Phaedrus, 265e*. 3̃70 BCE.

[195] Chris P. Ponting. The human cell atlas: Making ‘cell space’ for disease. *DMM Disease Models and Mechanisms*, 2019.

[196] Karl Popper. *Conjectures and refutations: The growth of scientific knowledge*. Routledge, 1963.

[197] Scott Powers, Matt DeJongh, Aaron A Best, and Nathan L Tintle. Cautions about the reliability of pairwise gene correlations based on expression data. *Frontiers in microbiology*, 6:650, 2015.

[198] SV Precious, CM Kelly, AE Reddington, NN Vinh, RC Stickland, V Pekarik, C Scherf, R Jeyasingham, J Glasbey, M Holeiter, et al. Foxp1 marks medium spiny neurons from precursors to maturity and is required for their differentiation. *Experimental neurology*, 282:9–18, 2016.

[199] Jason S Presnell, Christine E Schnitzler, and William E Browne. Klf/sp transcription factor family evolution: expansion, diversification, and innovation in eukaryotes. *Genome biology and evolution*, 7(8):2289–2309, 2015.

[200] Marcel Proust. *À la recherche du temps perdu*. Grasset and Gallimard, 1913–1927.

[201] Sidharth V Puram, Itay Tirosh, Anuraag S Parikh, Anoop P Patel, Keren Yizhak, Shawn Gillespie, Christopher Rodman, Christina L Luo, Edmund A Mroz, Kevin S Emerick, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell*, 171(7):1611–1624, 2017.

[202] Yuan Qi and Hui Ge. Modularity and dynamics of cellular networks. *PLoS computational biology*, 2(12):e174, 2006.

[203] Wenwei Qian, Zhiyuan Zhang, Wen Peng, Jie Li, Qiou Gu, Dongjian Ji, Qingyuan Wang, Yue Zhang, Bing Ji, Sen Wang, et al. Cdca3 mediates p21-dependent proliferation by regulating e2f1 expression in colorectal cancer. *International Journal of Oncology*, 53(5):2021–2033, 2018.

[204] Peng Qiu. Embracing the dropouts in single-cell RNA-seq analysis. *Nature Communications*, 2020.

[205] Jeffrey J Quinn and Howard Y Chang. Unique features of long non-coding rna biogenesis and function. *Nature Reviews Genetics*, 17(1):47–62, 2016.

[206] H. Rackham, W.H.S. Jones, and D.E. Eichholz. *Pliny: Natural History*. Number v. 2 in Pliny: Natural History. Harvard University Press, 1989.

[207] Pasko Rakic. Mode of cell migration to the superficial layers of fetal monkey neocortex. *Journal of Comparative Neurology*, 145(1):61–83, 1972.

[208] Pasko Rakic. Elusive radial glial cells: historical and evolutionary perspective. *Glia*, 43(1):19–32, 2003.

[209] Joseph Ramsey, Jiji Zhang, and Peter L Spirtes. Adjacency-faithfulness and conservative causal inference. *arXiv preprint arXiv:1206.6843*, 2012.

[210] Di Ran, Shanshan Zhang, Nicholas Lytal, and Lingling An. scdoc: correcting drop-out events in single-cell rna-seq data. *Bioinformatics*, 36(15):4233–4239, 2020.

[211] Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research*, 47(W1):W191–W198, 2019.

[212] Pavithran T Ravindran, Maxwell Z Wilson, Siddhartha G Jena, and Jared E Toettcher. Engineering combinatorial and dynamic decoders using synthetic immediate-early genes. *Communications biology*, 3(1):1–10, 2020.

[213] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. Science forum: the human cell atlas. *elife*, 6:e27041, 2017.

[214] P. Reps and N. Senzaki. *Zen Flesh, Zen Bones Classic Edition: A Collection of Zen and Pre-Zen Writings*. Tuttle Publishing, 2008.

[215] Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of artificial intelligence research*, 11:95–130, 1999.

[216] Regina H Reynolds, Juan Botía, Mike A Nalls, John Hardy, Sarah A Gagliano Taliun, and Mina Ryten. Moving beyond neurons: the role of cell type-specific gene regulation in parkinson's disease heritability. *NPJ Parkinson's disease*, 5(1):1–14, 2019.

[217] John A Rice. *Mathematical statistics and data analysis*. Cengage Learning, 2006.

[218] Robert W Robinson. Counting labeled acyclic digraphs. *New directions in the theory of graphs*, pages 239–273, 1973.

[219] Igor Rodchenkov, Ozgun Babur, Augustin Luna, Bulent Arman Aksoy, Jeffrey V Wong, Dylan Fong, Max Franz, Metin Can Siper, Manfred Cheung, Michael Wrana, et al. Pathway commons 2019 update: integration, analysis and exploration of pathway data. *Nucleic acids research*, 48(D1):D489–D497, 2020.

[220] Simone Romano, Nguyen Xuan Vinh, James Bailey, and Karin Verspoor. Adjusting for chance clustering comparison measures. *The Journal of Machine Learning Research*, 17(1):4635–4666, 2016.

[221] Luciana Ferreira Romão, Vivian de Oliveira Sousa, Vivaldo Moura Neto, and Flávia Carvalho Alcantara Gomes. Glutamate activates gfap gene promoter from cultured astrocytes through tgf-β1 pathways. *Journal of neurochemistry*, 106(2):746–756, 2008.

[222] Fernando E Rosas, Pedro AM Mediano, Andrea I Luppi, Thomas F Varley, Joseph T Lizier, Sebastiano Stramaglia, Henrik J Jensen, and Daniele Marinazzo. Disentangling high-order mechanisms and high-order behaviours in complex systems. *Nature Physics*, pages 1–2, 2022.

[223] Kenneth M Rosen, Matthew A McCormack, Lydia Villa-Komaroff, and George D Mower. Brief visual experience induces immediate early gene expression in the cat visual cortex. *Proceedings of the National Academy of Sciences*, 89(12):5437–5441, 1992.

[224] Gian-Carlo Rota. On the foundations of combinatorial theory i. theory of möbius functions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 2(4):340–368, 1964.

[225] Tammo Rukat, Chris C. Holmes, Michalis K. Titsias, and Christopher Yau. Bayesian boolean matrix factorisation. In *34th International Conference on Machine Learning, ICML 2017*, 2017.

[226] Marilyn Safran, Naomi Rosen, Michal Twik, Ruth BarShir, Tsippi Iny Stein, Dvir Dahary, Simon Fishilevich, and Doron Lancet. The genecards suite. In *Practical guide to life science databases*, pages 27–56. Springer, 2021.

[227] Ansuman T Satpathy, Jeffrey M Granja, Kathryn E Yost, Yanyan Qi, Francesca Meschi, Geoffrey P McDermott, Brett N Olsen, Maxwell R Mumbach, Sarah E Pierce, M Ryan Corces, et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral t cell exhaustion. *Nature biotechnology*, 37(8):925–936, 2019.

[228] Martin H Schaefer, Luis Serrano, and Miguel A Andrade-Navarro. Correcting for the study bias associated with protein–protein interaction measurements reveals differences between protein degree distributions from different cancer types. *Frontiers in genetics*, 6:260, 2015.

[229] Michel Schmitt-Ney. The foxo's advantages of being a family: considerations on function and evolution. *Cells*, 9(3):787, 2020.

[230] Matthew T Schmitz, Kadellyn Sandoval, Christopher P Chen, Mohammed A Mostajo-Radji, William W Seeley, Tomasz J Nowakowski, Chun Jimmie Ye, Mercedes F Paredes, and Alex A Pollen. The development and evolution of inhibitory neurons in primate cerebrum. *Nature*, 603(7903):871–877, 2022.

[231] A Kristin Schneider, Giuseppe Cama, Mandeep Ghuman, Francis J Hughes, and Borzo Gharibi. Sprouty 2, an early response gene regulator of fosb and mesenchymal stem cell proliferation during mechanical loading and osteogenic differentiation. *Journal of Cellular Biochemistry*, 118(9):2606–2614, 2017.

[232] Erwin M Schoof, Benjamin Furtwängler, Nil Üresin, Nicolas Rapin, Simonas Savickas, Coline Gentil, Eric Lechman, John E Dick, Bo T Porse, et al. Quantitative single-cell proteomics as a tool to characterize cellular hierarchies. *Nature communications*, 12(1):1–15, 2021.

[233] David Schwab, Ilya Nemenman, and Pankaj Mehta. Zipf's law and criticality in multivariate data without fine-tuning. *Physical review letters*, 113, 10 2013.

[234] Julian D Schwab, Silke D Kühlwein, Nensi Ikonomi, Michael Kühl, and Hans A Kestler. Concepts in boolean network modeling: What do they all mean? *Computational and structural biotechnology journal*, 18:571–582, 2020.

[235] Arnau Sebé-Pedrós, Baptiste Saudemont, Elad Chomsky, Flora Plessier, Marie-Pierre Mailhé, Justine Renno, Yann Loe-Mie, Aviezer Lifshitz, Zohar Mukamel, Sandrine Schmutz, et al. Cnidarian cell type diversity and regulation revealed by whole-organism single-cell rna-seq. *Cell*, 173(6):1520–1534, 2018.

[236] Danielle Seurin, Alain Lombet, Sylvie Babajko, Francois Godeau, and Jean-Marc Ricort. Insulin-like growth factor binding proteins increase intracellular calcium levels in two different cell lines. *PloS one*, 8(3):e59323, 2013.

[237] M. Sheldrake. *Entangled Life: How Fungi Make Our Worlds, Change Our Minds & Shape Our Futures*. Random House Publishing Group, 2021.

[238] Hui Z Sheng, Peng X Lin, and Phillip G Nelson. Combinatorial expression of immediate early genes in single neurons. *Molecular brain research*, 30(2):196–202, 1995.

[239] Morgan Sheng and Michael E Greenberg. The regulation and function of c-fos and other immediate early genes in the nervous system. *Neuron*, 4(4):477–485, 1990.

[240] Yingchao Shi, Mengdi Wang, Da Mi, Tian Lu, Bosong Wang, Hao Dong, Suijuan Zhong, Youqiao Chen, Le Sun, Xin Zhou, et al. Mouse and human share conserved transcriptional programs for interneuron development. *Science*, 374(6573):eabj6641, 2021.

[241] Hidetoshi Shimodaira. An approximately unbiased test of phylogenetic tree selection. *Systematic biology*, 51(3):492–508, 2002.

[242] Shaleen Shrestha, Jared Allan Sewell, Clarissa Stephanie Santoso, Elena Forchielli, Sebastian Carrasco Pro, Melissa Martinez, and Juan Ignacio Fuxman Bass. Discovering human transcription factor physical interactions with genetic variants, novel dna motifs, and repetitive elements using enhanced yeast one-hybrid assays. *Genome research*, 29(9):1533–1544, 2019.

[243] DL Simmons, BG Neel, R Stevens, G Evett, and RL Erikson. Identification of an early-growth-response gene encoding a novel putative protein kinase. *Molecular and cellular biology*, 12(9):4164–4169, 1992.

[244] Per Sebastian Skardal and Alex Arenas. Higher order interactions in complex networks of phase oscillators promote abrupt synchronization switching. *Communications Physics*, 3(1):1–6, 2020.

[245] N.J.A. Sloane. The on-line encyclopedia of integer sequences, entry a003024, 2022.

[246] I Smart. The subependymal layer of the mouse brain and its cell production as shown by radioautography after thymidine-h3 injection. *Journal of Comparative Neurology*, 116(3):325–347, 1961.

[247] MARION EDMONDS SMITH. The metabolism of myelin lipids. *Advances in lipid research*, 5:241–278, 1967.

[248] Stephen J Smith, Uygar Sümbül, Lucas T Graybuck, Forrest Collman, Sharmishtaa Seshamani, Rohan Gala, Olga Gliko, Leila Elabbady, Jeremy A Miller, Trygve E Bakken, et al. Single-cell transcriptomic evidence for dense intracortical neuropeptide networks. *Elife*, 8:e47889, 2019.

[249] Xiaolei Song, Haotian Chen, Zicong Shang, Heng Du, Zhenmeiyu Li, Yan Wen, Guoping Liu, Dashi Qi, Yan You, Zhengang Yang, et al. Homeobox gene six3 is required for the differentiation of d2-type medium spiny neurons. *Neuroscience Bulletin*, 37(7):985–998, 2021.

[250] You-Hyang Song, Jiwon Yoon, and Seung-Hee Lee. The role of neuropeptide somatostatin in the brain and its application in treating neurological disorders. *Experimental & Molecular Medicine*, 53(3):328–338, 2021.

[251] Martino Sorbaro, J Michael Herrmann, and Matthias Hennig. Statistical models of neural activity, criticality, and zipf's law. *The functional role of critical dynamics in neural systems*, pages 265–287, 2019.

[252] Matthew L Speir, Aparna Bhaduri, Nikolay S Markov, Pablo Moreno, Tomasz J Nowakowski, Irene Papatheodorou, Alex A Pollen, Brian J Raney, Lucas Seninge, W James Kent, et al. Ucsc cell browser: visualize your single-cell data. *Bioinformatics*, 37(23):4578–4580, 2021.

[253] Geoffrey Stanley, Ozgun Gokce, Robert C Malenka, Thomas C Südhof, and Stephen R Quake. Continuous and discrete neuron types of the adult murine striatum. *Neuron*, 105(4):688–699, 2020.

[254] Richard P Stanley. Enumerative combinatorics volume 1 second edition. *Cambridge studies in advanced mathematics*, 2011.

[255] Malayannan Subramaniam, John R Hawse, Nalini M Rajamannan, James N Ingle, and Thomas C Spelsberg. Functional role of klf10 in multiple disease processes. *Biofactors*, 36(1):8–18, 2010.

[256] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[257] Motoyosi Sugita. Functional analysis of chemical systems in vivo using a logical circuit equivalent. *J. theor. Biol*, 1(4):415, 1961.

[258] Beata Surmacz, Parinya Noisa, Jessica R Risner-Janiczek, Kailyn Hui, Mark Ungless, Wei Cui, and Meng Li. Dlk1 promotes neurogenesis of human and mouse pluripotent stem cell-derived neural progenitors via modulating notch and bmp signalling. *Stem Cell Reviews and Reports*, 8(2):459–471, 2012.

[259] Polina Suter, Jack Kuipers, Giusi Moffa, and Niko Beerenwinkel. Bayesian structure learning and sampling of bayesian networks with the r package bidag. *arXiv preprint arXiv:2105.00488*, 2021.

[260] Ryota Suzuki and Hidetoshi Shimodaira. Pvclust: an r package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12):1540–1542, 2006.

[261] Pamela J Swiatek and Thomas Gridley. Perinatal lethality and defects in hindbrain development in mice homozygous for a targeted mutation of the zinc finger gene krox20. *Genes & development*, 7(11):2071–2084, 1993.

[262] Gábor J. Székely and Maria L. Rizzo. Partial distance correlation with methods for dissimilarities. *The Annals of Statistics*, 42(6):2382 – 2412, 2014.

[263] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, Dec 2007.

[264] Damian Szklarczyk, Annika L Gable, Katerina C Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer

Bork, et al. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612, 2021.

[265] K. Tanahashi. *Treasury of the True Dharma Eye: Zen Master Dogen's Shobo Genzo*. Shambhala, 2013.

[266] Amos Tanay, Aviv Regev, and Ron Shamir. Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *Proceedings of the National Academy of Sciences*, 102(20):7203–7208, 2005.

[267] Bosiljka Tasic, Vilas Menon, Thuc Nghi Nguyen, Tae Kyung Kim, Tim Jarsky, Zizhen Yao, Boaz Levi, Lucas T Gray, Staci A Sorensen, Tim Dolbeare, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature neuroscience*, 19(2):335–346, 2016.

[268] Vladimir B Teif. Predicting gene-regulation functions: lessons from temperate bacteriophages. *Biophysical journal*, 98(7):1247–1256, 2010.

[269] René Thomas. Boolean formalization of genetic control circuits. *Journal of theoretical biology*, 42(3):563–585, 1973.

[270] Neha Tiwari, Abhijeet Pataskar, Sophie Péron, Sudhir Thakurela, Sanjeeb Kumar Sahu, María Figueres-Oñate, Nicolás Marichal, Laura López-Mascaraque, Vijay K. Tiwari, and Benedikt Berninger. Stage-specific transcription factors drive astrogliogenesis by remodeling gene regulatory landscapes. *Cell Stem Cell*, 23(4):557 – 571.e8, 2018.

[271] Gasper Tkacik, Elad Schneidman, Michael J Berry II, and William Bialek. Ising models for networks of real neurons. *arXiv preprint q-bio/0611072*, 2006.

[272] Lauren J Tracey, Yeji An, and Monica J Justice. Cytof: An emerging technology for single-cell proteomics in the mouse. *Current Protocols*, 1(4):e118, 2021.

[273] Cole Trapnell. Defining cell types and states with single-cell genomics. *Genome research*, 25(10):1491–1498, 2015.

[274] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nature biotechnology*, 32(4):381, 2014.

[275] John W Tullai, Michael E Schaffer, Steven Mullenbrock, Gabriel Sholder, Simon Kasif, and Geoffrey M Cooper. Immediate-early and delayed primary response genes are distinct in function and genomic architecture. *Journal of Biological Chemistry*, 282(33):23981–23995, 2007.

[276] Aleksandra Turanjanin. Stability of hierarchical clustering. 2020.

[277] Kelsey M Tyssowski, Nicholas R DeStefino, Jin-Hyung Cho, Carissa J Dunn, Robert G Poston, Crista E Carty, Richard D Jones, Sarah M Chang, Palmyra Romeo, Mary K Wurzelmann, et al. Different neuronal activity patterns induce different gene expression programs. *Neuron*, 98(3):530–546, 2018.

[278] Mathias Uhlén, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, et al. Proteomics. tissue-based map of the human proteome. *Science (New York, NY)*, 347(6220):1260419–1260419, 2015.

[279] Annalaura Vacca, Masayoshi Itoh, Hideya Kawaji, Erik Arner, Timo Lassmann, Carsten O Daub, Piero Carninci, Alistair RR Forrest, Yoshihide Hayashizaki, FANTOM Consortium, et al. Conserved temporal ordering of promoter activation implicates common mechanisms governing the immediate early response across cell types and stimuli. *Open biology*, 8(8):180011, 2018.

[280] Susanne C van den Brink, Fanny Sage, Ábel Vértesy, Bastiaan Spanjaard, Josi Peterson-Maduro, Chloé S Baron, Catherine Robin, and Alexander Van Oudenaarden. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nature methods*, 14(10):935–936, 2017.

[281] David Van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716–729, 2018.

[282] Keith W Vance and Chris P Ponting. Transcriptional regulatory functions of nuclear long noncoding rnas. *Trends in Genetics*, 30(8):348–355, 2014.

[283] Thomas Varley and Erik Hoel. Emergence as the conversion of information: A unifying theory. pages 1–20, 2021.

[284] Luke F Vistain and Savaş Tay. Single-cell proteomics. *Trends in Biochemical Sciences*, 46(8):661–672, 2021.

[285] Christine Vogel and Edward M Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews genetics*, 13(4):227–232, 2012.

[286] Alexander Von Humboldt. *Kosmos: Entwurf einer physischen Weltbeschreibung*, volume 1. FW Thomas, 1845.

[287] Günter P Wagner, Mihaela Pavlicev, and James M Cheverud. The road to modularity. *Nature Reviews Genetics*, 8(12):921–931, 2007.

[288] Michael Wainberg, Roarke A. Kamber, Akshay Balsubramani, Robin M. Meyers, Nasa Sinnott-Armstrong, Daniel Hornburg, Lihua Jiang, Joanne Chan, Ruiqi Jian, Mingxin Gu, Anna Shcherbina, Michael M. Dubreuil, Kaitlyn Spees, Wouter Meuleman, Michael P. Snyder, Michael C. Bassik, and Anshul Kundaje. A genome-wide atlas of co-essential modules assigns function to uncharacterized genes. *Nature Genetics*, 2021.

[289] James Z Wang, Zhidian Du, Rapeeporn Payattakool, Philip S Yu, and Chin-Fu Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281, 2007.

[290] Lu Wang, Tze King Tan, Adam D Durbin, Mark W Zimmerman, Brian J Abraham, Shi Hao Tan, Phuong Cao Thi Ngoc, Nina Weichert-Leahey, Koshi Akahane,

Lee N Lawton, et al. Ascl1 is a mycn-and lmo1-dependent member of the adrenergic neuroblastoma core regulatory circuitry. *Nature communications*, 10(1):1–15, 2019.

[291] Yi Wang, Stephanie C Hicks, and Kasper D Hansen. Addressing the mean-correlation relationship in co-expression analysis. *PLoS computational biology*, 18(3):e1009954, 2022.

[292] John Watkinson, Kuo Ching Liang, Xiadong Wang, Tian Zheng, and Dimitris Anastassiou. Inference of regulatory gene interactions from expression data using three-way mutual information. *Annals of the New York Academy of Sciences*, 2009.

[293] Andreas PM Weber. Discovering new biology through sequencing of rna. *Plant physiology*, 169(3):1524–1531, 2015.

[294] Caleb Weinreb, Alejo Rodriguez-Fraticelli, Fernando D Camargo, and Allon M Klein. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science*, 367(6479):eaaw3381, 2020.

[295] Caleb Weinreb, Samuel Wolock, Betsabeh K Tusi, Merav Socolovsky, and Allon M Klein. Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences*, 115(10):E2467–E2476, 2018.

[296] S. Werhausen. Magneto. https://github.com/s9w/magneto, 2015.

[297] Michael L Whitfield, Gavin Sherlock, Alok J Saldanha, John I Murray, Catherine A Ball, Karen E Alexander, John C Matese, Charles M Perou, Myra M Hurt, Patrick O Brown, et al. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular biology of the cell*, 13(6):1977–2000, 2002.

[298] Norbert Wiener. Cybernetics: Control and communication in the animal and the machine–2nd. 1961.

[299] Minde Willardsen, David A Hutcheson, Kathryn B Moore, and Monica L Vetter. The ets transcription factor etv1 mediates fgf signaling to initiate proneural gene expression during xenopus laevis retinal development. *Mechanisms of development*, 131:57–67, 2014.

[300] Paul L. Williams and Randall D. Beer. Nonnegative Decomposition of Multivariate Information. pages 1–14, 2010.

[301] C Winkler and S Yao. The midkine family of growth factors: diverse roles in nervous system formation and maintenance. *British journal of pharmacology*, 171(4):905–912, 2014.

[302] Samuel L Wolock, Romain Lopez, and Allon M Klein. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell systems*, 8(4):281–291, 2019.

[303] DD Wood, GJ Vella, and MA Moscarello. Interaction between human myelin basic protein and lipophilin. *Neurochemical research*, 9(10):1523–1531, 1984.

[304] Andrea Wulf. *The Invention of Nature: The Adventures of Alexander von Humboldt, the Lost Hero of Science: Costa & Royal Society Prize Winner*. Hachette UK, 2015.

[305] Michelle L Wynn, Nikita Consul, Sofia D Merajver, and Santiago Schnell. Logic-based models in systems biology: a predictive and parameter-free network analysis method. *Integrative biology*, 4(11):1323–1337, 2012.

[306] Peng-Fei Xia, Hua Ling, Jee Loon Foo, and Matthew Wook Chang. Synthetic genetic circuits for programmable biological functionalities. *Biotechnology Advances*, 37(6):107393, 2019.

[307] Xiong Xiao, Hanfei Deng, Alessandro Furlan, Tao Yang, Xian Zhang, Ga-Ram Hwang, Jason Tucciarone, Priscilla Wu, Miao He, Ramesh Palaniswamy, et al. A genetically defined compartmentalized striatal direct pathway for negative reinforcement. *Cell*, 183(1):211–227, 2020.

[308] Zhejun Xu, Qifei Liang, Xiaolei Song, Zhuangzhi Zhang, Susan Lindtner, Zhenmeiyu Li, Yan Wen, Guoping Liu, Teng Guo, Dashi Qi, et al. Sp8 and sp9 coordinately promote d2-type medium spiny neuron production by activating six3 expression. *Development*, 145(14):dev165456, 2018.

[309] Yoshihide Yamaguchi, Kazuhiro Ikenaka, Michio Niinobe, Hitoshi Yamada, and Katsuhiko Mikoshiba. Myelin proteolipid protein (plp), but not dm-20, is an inositol hexakisphosphate-binding protein. *Journal of Biological Chemistry*, 271(44):27838–27846, 1996.

[310] Ee-Lynn Yap and Michael E Greenberg. Activity-regulated transcription: bridging the gap between neural activity and behavior. *Neuron*, 100(2):330–348, 2018.

[311] Jejoong Yoo, Hajin Kim, Aleksei Aksimentiev, and Taekjip Ha. Direct evidence for sequence-dependent attraction between double-stranded dna controlled by methylation. *Nature communications*, 7(1):1–7, 2016.

[312] Guangchuang Yu, Fei Li, Yide Qin, Xiaochen Bo, Yibo Wu, and Shengqi Wang. Gosemsim: an r package for measuring semantic similarity among go terms and gene products. *Bioinformatics*, 26(7):976–978, 2010.

[313] Amit Zeisel, Hannah Hochgerner, Peter Lönnerberg, Anna Johnsson, Fatima Memic, Job van der Zwan, Martin Häring, Emelie Braun, Lars E. Borm, Gioele La Manno, Simone Codeluppi, Alessandro Furlan, Kawai Lee, Nathan Skene, Kenneth D. Harris, Jens Hjerling-Leffler, Ernest Arenas, Patrik Ernfors, Ulrika Marklund, and Sten Linnarsson. Molecular Architecture of the Mouse Nervous System. *Cell*, 2018.

[314] David Zemmour, Rapolas Zilionis, Evgeny Kiner, Allon M Klein, Diane Mathis, and Christophe Benoist. Single-cell gene expression reveals a landscape of regulatory t cell phenotypes shaped by the tcr. *Nature immunology*, 19(3):291–301, 2018.

[315] Allen W Zhang, Ciara O'Flanagan, Elizabeth A Chavez, Jamie LP Lim, Nicholas Ceglia, Andrew McPherson, Matt Wiens, Pascale Walters, Tim Chan, Brittany

Hewitson, et al. Probabilistic cell-type assignment of single-cell rna-seq for tumor microenvironment profiling. *Nature methods*, 16(10):1007–1015, 2019.

[316] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[317] Yilan Zhang, Yuqun Cai, Yafei Wang, Xin Deng, Yifan Zhao, Yubin Zhang, and Yunli Xie. Survival control of oligodendrocyte progenitor cells requires the transcription factor 4 during olfactory bulb development. *Cell death & disease*, 12(1):1–14, 2021.

[318] Ze Zhang, Danni Luo, Xue Zhong, Jin Huk Choi, Yuanqing Ma, Stacy Wang, Elena Mahrt, Wei Guo, Eric W Stawiski, Zora Modrusan, et al. Scina: a semi-supervised subtyping algorithm of single cells and bulk samples. *Genes*, 10(7):531, 2019.

[319] Zhuangzhi Zhang, Zicong Shang, Lin Yang, Ziwu Wang, Yu Tian, Yanjing Gao, Zihao Su, Rongliang Guo, Weiwei Li, Guoping Liu, et al. The transcription factor zfp503 promotes the d1 msn identity and represses the d2 msn identity. 2022.

[320] Zheng-dong Zhao, Xiao Han, Renchao Chen, Yiqiong Liu, Aritra Bhattacherjee, Wenqiang Chen, and Yi Zhang. A molecularly defined d1 medium spiny neuron subtype negatively regulates cocaine addiction. *Science advances*, 8(31):eabn3552, 2022.

> And every science, when we understand it not as an instrument of power and domination but as an adventure in knowledge pursued by our species across the ages, is nothing but this harmony, more or less vast, more or less rich from one epoch to another, which unfurls over the course of generations and centuries, by the delicate counterpoint of all the themes appearing in turn, as if summoned from the void.
>
> Alexander Grothendieck [98, 121]